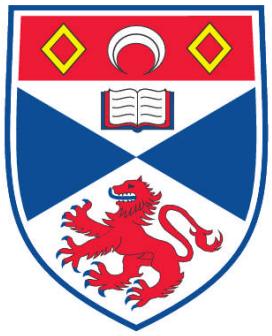
DOWN TO EARTH PHILOSOPHY: AN ANTI-EXCEPTIONALIST ESSAY ON THOUGHT EXPERIMENTS AND PHILOSOPHICAL METHODOLOGY

Daniele Sgaravatti

A Thesis Submitted for the Degree of PhD
at the
University of St. Andrews



2012

DOWN TO EARTH PHILOSOPHY

AN ANTI-EXCEPTIONALIST ESSAY ON THOUGHT EXPERIMENTS AND PHILOSOPHICAL METHODOLOGY

DANIELE SGARAVATTI

Down to Earth Philosophy

An Anti-exceptionalist Essay on Thought Experiments and Philosophical Methodology

Daniele Sgaravatti

Submitted for the degree of Doctor in Philosophy

08/06/2011

Abstract

In the first part of the dissertations, chapters 1 to 3, I criticize several views which tend to set philosophy apart from other cognitive achievements. I argue against the popular views that 1) Intuitions, as a *sui generis* mental state, are involved crucially in philosophical methodology 2) Philosophy requires engagement in conceptual analysis, understood as the activity of considering thought experiments with the aim to throw light on the nature of our concepts, and 3) Much philosophical knowledge is a priori. I do not claim to have a proof that nothing in the vicinity of these views is correct; such a proof might well be impossible to give. However, I consider several versions, usually prominent ones, of each of the views, and I show those versions to be defective. Quite often, moreover, different versions of the same worry apply to different versions of the same theory.

In the fourth chapter I discuss the epistemology of the judgements involved in philosophical thought experiments, arguing that their justification depends on their being the product of a competence in applying the concepts involved, a competence which goes beyond the possession of the concepts. I then offer, drawing from empirical psychology, a sketch of the form this cognitive competence could take. The overall picture squares well with the conclusions of the first part. In the last chapter I consider a challenge to the use of thought experiments in contemporary analytic philosophy coming from the 'experimental philosophy' movement. I argue that there is no way of individuating the class of hypothetical judgements under discussion which makes the challenge both interesting and sound. Moreover, I argue that there are reasons to think that philosophers possess some sort of expertise which sets them apart from non-philosophers in relevant ways.

I, Daniele Sgaravatti, hereby certify that this thesis, which is approximately 68000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September 2007 and as a candidate for the degree of Doctor of Philosophy in September 2007; the higher study for which this is a record was carried out in the University of St Andrews between 2007 and 2011.

Date: 08/06/2011 Signature of candidate ……

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 08/06/2011 Signature of supervisor ……

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis: Access to printed copy and electronic publication of thesis through the University of St Andrews.

Signature of candidate …… Signature of supervisor ……

For my mother,

who taught me not to change my mind too easily

Table of Contents

Acknowledgements

First of all, I need to thank Jessica Brown, who has been my main supervisor at St Andrews for almost four years. She always helped me to make the best of the thoughts I had, by showing me ways in which they could be challenged and ways in which they could be developed, and also by helping me to just express them in the most comprehensible and effective ways. I believe I have been extremely lucky to have the opportunity to work with her, and I hope that some of her capacity to combine a grasp of the bigger picture with attention to the details will stay with me.

Brian Weatherson has been my secondary supervisor in the last few years. He also helped me greatly with the work of this dissertation. He was always extremely generous with his time, and discussing philosophy with him has always been a pleasure as well as a source of philosophical insight.

Many other people read parts of this dissertation and provided me with invaluable feedback. These are Simona Aimar, Derek Ball, Andrea Bianchi, Herman Cappelen, Dave Chalmers, Ole Koksvik, Julia Langkau, Daria Mingardo and Ernest Sosa.

I discussed the issues of this work with too many people for me to even attempt at mentioning everyone who gave me useful suggestions; I apologize to everyone who is not mentioned. Still, I wish to thank, for having had the patience to discuss the topics of this work with me, John Bengson, Björn Brodowski, Yuri Cath, Tamar Szabó Gendler, Torfinn Huvenes, Andrea Onofri, Jonathan Schaffer, Anders Schoubye, Max Seeger, Martin Smith, Andreas Stokke, Tim Williamson, Elia Zardini, and my colleagues at the Emmy Noether group at the University of Cologne; in particular Brendan and Magdalena Balcerak-Jackson, Lars Dänzer and Joachim Horvath. While I was working at this dissertation I also spent time at the department of philosophy at Rutgers, and at the CSMN research centre in Oslo. In both places I felt at home, and I wish to thank everyone who made those visits possible and everyone who made them so pleasant. But of course my main home for many years has been St Andrews. I really need to thank the Arché research centre and the department of St Andrews as a whole, for providing such a wonderful environment for doing philosophy, and for building friendships.

Finally, I would like to thank again my wife, Daria Mingardo, for being there.

Introduction

This is, as the title suggests, in large part an essay in philosophical methodology. Before introducing its content, I would like to say something about the interest of philosophical methodology. I have been asked fairly often, by philosophers, in a slightly aggressive way, the following: What is so interesting about philosophical methodology? I always find it slightly surprising that a philosopher would ask that question. In general, I would expect someone who pursues an activity to have some interest in thinking about the way that activity is, or should be, conducted. Philosophers, in particular, are usually interested in reflections upon almost any subject matter; it seems bizarre that, of all the things in the universe, the one they do not want to think about is philosophical methodology. This lack of interest is even more surprising because philosophical methodology is not really different from (a part of) philosophy. I hope this will be clear by looking at the topics discussed in this work too. The questions I will address are mostly questions in philosophy of language, philosophy of mind and epistemology.

I have come to think that there are historical reasons why philosophical methodology is viewed with suspicion, and I will offer some very brief speculations. Most great philosophers of the western tradition were concerned with questions of method in philosophy, although they often did not distinguish between the method of philosophy and the method of the sciences, or even the method of the search for knowledge in general. However, for analytic philosophers at least, the tendency to scrutinize philosophical methods is mostly associated with the neo-positivists. I am not mainly interested here in historical accuracy; the way the history of analytic philosophy is perceived is more relevant than the actual history in this context. So, let us say that the neo-positivists thought that the history of philosophy was largely a history of failures, due to the lack of a proper methodology; and that a proper methodology, involving the use of formal languages, would lead to a relatively easy solution for every philosophical problem that could be solved, and to the abandonment of those problems that could not be solved as lacking in content. The emphasis on clarity, precision and rigour that accompanied this view of philosophical methodology is of course at the heart

of the self-conception of analytic philosophy. There is even something to be said, I believe, for the iconoclastic rejection of part of the western philosophical tradition. It might be naïve, but it might also give one the freedom to start afresh when it is needed; and after all, some great philosophers of the past, such as Descartes, had a similar attitude. However, at the same time, the idea that there is such a thing as a proper philosophical method, which can be stated in formal terms, and which will allow for a solution of every problem, is now justly perceived as a naive dream. Moreover, the idea that every investigation outside the proper method is to be denounced as nonsense is justly perceived as a dangerous form of dogmatism.

I should immediately dissipate the suspicion that, in taking philosophical methodology seriously, I wish to bring back the neo-positivist strictures. It might be useful to distinguish between a prescriptive and a descriptive form of philosophical methodology. The neo-positivists were after a prescriptive sort of methodology, a set of general rules which could, and should, be self-consciously employed by philosophers, changing radically the way the discipline was practiced. There was indeed a renewed interest in philosophical methodology in the last twenty years or so, but it is, mostly[1], an interest in describing and understanding the way philosophy is practiced. This we might call descriptive philosophical methodology. In the sense in which I am using it here, 'descriptive' is not to be contrasted with 'normative'. One can describe using a normative vocabulary (and thereby give an *indirect* prescription). General epistemology provides a good example in this sense: epistemologists try to specify the conditions under which a belief is justified or constitutes knowledge. However, for the most part, they do not expect that subjects can usefully be taught what the conditions for justified belief or knowledge are, or that subjects consciously try to meet them. I recommend a similar attitude in philosophical methodology. Philosophers, like all subjects, should in my view attempt to solve the problems they have at hand being guided primarily by the content of those problems, and no prescription much more specific than 'do your best!' is needed, or useful. I am not saying that we are guaranteed to be already using good methods, or even guaranteed not to fall into talking nonsense. I am only saying that there is no general method to choose good methods that we can mechanically apply, and nothing but intellectual honesty (and perhaps some good luck) can save one from working in a degenerate research program, in philosophy as elsewhere.

However, it should be clear that there is intrinsic interest in a better understanding of philosophical methodology, in the descriptive sense. Philosophers, as I said at the outset, are

[1] In this respect the experimental philosophy movement, which will be discussed in chapter 5, is an exception, at least in some of its aspects.

supposed to be curious about the nature of the world, and the nature of human activities in particular; it would be really peculiar if such curiosity did not extend to the human activity of philosophy itself. This work will explore several interconnected issues in this area. The general approach defended here is 'anti-exceptionalist'; I indicate with this term the rejection of 'philosophical exceptionalism', the idea that philosophical method is exceptional, in the sense of being radically different from the method employed in all, or most, other disciplines (the expression 'philosophical exceptionalism' was first used, as far as I know, in Williamson [2007], who also contrasts his view with it). This is of course a vague definition. Philosophical exceptionalism is a sort of attitude, more than a precise set of views, and so is anti-exceptionalism. Of course the discussion to follow will focus on precise claims, giving a more definite content to those attitudes.

It will be helpful at this point to say something about the relation between anti-exceptionalism and another famous 'ism' of philosophical methodology, namely naturalism. Even focusing on *methodological* naturalism (thereby setting aside the sense of 'naturalism' in which it is identical to physicalism or some other ontological thesis), naturalism can be understood in several different ways; there are two in particular, which are clearly connected to the etymology of the term, that I would like to contrast with anti-exceptionalism. On one hand, there is the idea that we should not appeal to super-natural powers or entities to explain facts. Although of course it is debatable what counts as super-natural, I take it that the thesis is trivially correct on any reasonable interpretation. So I wish to set it aside as lacking interest. On the other hand, there is the much stronger idea that the proper method of investigating the world is the method of the *natural* sciences. This is indeed an interesting idea, but it is not one I wish to endorse. I see philosophy as continuous with other sciences, as well as with common sense; in this respect, anti-exceptionalism is a Quinean thesis. But the other sciences that philosophy is continuous to include not only physics or biology, but also disciplines as diverse as history, psychology, linguistics and, crucially, mathematics (the point was also recently noted in Williamson [2011b]); the latter cannot be counted as a *natural* science even stretching the term. We cannot say that these disciplines all use the method of natural sciences like physics and biology, unless we give a description of such a method too vague to be of any use (indeed, I find it dubious that there is a useful sense in which physics and biology use the same method). A closely connected issue is whether we can usefully describe the method of philosophy as empirical; I favor a negative answer (although I also deny that the method is a priori), as I explain briefly at the start of chapter 3, where I come back to the matter of the

relation between my criticisms of the a priori/a posteriori distinction and those in the directly Quinean tradition.

Here is the plan of the work. The first three chapters comprise a negative part of the work. They argue against some strong claims that have been made about philosophical methodology. The chapters address those claims in order of increasing popularity. In addition, the arguments in chapter 3 rest in part on the conclusions of the first two chapters. A first claim, discussed in chapter 1, is that philosophy makes a central use of something called 'intuitions'. An intuition, in some versions of this view at least, is a mental state that somehow puts us in contact with abstract matters. Typically, according to this view, one can use one's own intuitions to determine what it would be correct to say about certain hypothetical scenarios. I do not contest that philosophy makes use of hypothetical scenarios, or thought experiments, and later I am going to develop a view about that use. But I will argue that we have no reason to appeal to the notion of intuition, in the sense of a *sui generis* mental state. I make room for a peculiar mental state that we sometimes experience when considering a thought experiment, and which we might call a 'seeming'. I argue that seemings can be identified with felt inclinations to believe, and that, although they can constitute evidence, there is no reason to believe that they play a foundational or otherwise central epistemic role. Their evidential role is limited, and derivative on that of the corresponding judgements.

The second chapter discusses a view of philosophical methodology which was, and to some extent still is, widely held in analytic philosophy. The view is that philosophers are engaged in the activity of conceptual analysis. Conceptual analysis, very roughly, is supposed to proceed as follows: we consider some imaginary cases and we judge whether certain concepts would be properly applied in those cases. The purpose of such an activity is to shed light on the concepts themselves. This is possible because the judgements involved are constitutively tied to the concepts in some way; they are *analytic*, in some sense of this term. In chapter 2, I argue that there is no conception of analyticity which will fit the bill. I rely on arguments presented in Williamson [2007], but I also focus on a defense of conceptual analysis which he did not consider explicitly, the epistemic two-dimensionalist framework developed by Frank Jackson and David Chalmers. I argue that the framework is fundamentally flawed by its inability to make room for the idea that mental and linguistic content is not internalistic, i.e. it does not depend exclusively on the intrinsic properties of the individual entertaining the content, but also depend on relations between the individual and the world which she or he inhabits.

The third chapter discusses the distinction between a priori and a posteriori knowledge. Some philosophers argue, or even take for granted, that philosophy provides a priori knowledge if it provides knowledge of any kind. I argue (expanding on arguments presented in Hawthorne [2007] and Williamson [2007]) that the distinction is not sufficiently clear to play any important theoretical role. A priori knowledge is supposed to be knowledge independent of experience. Yet, we lack a good grasp of this notion in turn. On some ways of spelling it out, too much ends up counting as a priori, and on other ways of spelling it out, too much ends up counting as a posteriori. Of course what is 'too much' might be a matter of disagreement, but I here take as a standard what has been traditionally thought of as the extension of the distinction: roughly, philosophy, logic and mathematics are a priori, natural sciences and (most) knowledge of contingent truths are a posteriori. I assume that if the distinction's extension ends up being radically different from this standard, the distinction cannot play the same theoretical role that it used to play. In particular, one cannot hold that philosophical thought experiments are a priori without classifying as a priori a number of apparently ordinary claims outside philosophy, logic, mathematics, or anything else that was traditionally considered a region in which a priori knowledge was possible. In the chapter, I consider three different ways of saving the distinction; the first, proposed by Carrie Jenkins (Jenkins [2008a], [2008b]), appeals to the notion of epistemic grounds. A second one appeals to the modal status of the content of the claims to be judged a priori or a posteriori. A third appeals to the notion of reasoning independent of experience. The first and the third proposal fail to even approximate the traditional extensions of the distinction. The second comes closer, but, besides a number of other problems, it fails to provide a genuine epistemological distinction.

Chapters 4 and 5 constitute the positive part of the work. One might legitimately wonder how an anti-exceptionalist, given the rejection of appeals to intuitions, analyticity and apriority, can positively characterize philosophical methodology (a similar worry pressed for example in Malmgren [forthcoming]). Chapter 4 looks in some detail at one tool of philosophical methodology which, as we have seen, is considered central by many defenders of philosophical exceptionalism, i.e. thought experiments. I start by reconstructing the logical structure of at least one use of thought experiments, the one in which they are involved in the construction of an argument. I argue (following Williamson [2007]) that the crucial step in reasoning with thought experiments (what is sometimes called the 'intuition' about the case) is the forming of a counterfactual judgement about what would be the case, were the hypothetical scenario to be actual. I discuss some objections to this view, in particular those

raised by Ichikawa and Jarvis [2009], and I show that they are unconvincing. I then turn to discuss the epistemic basis of our judgements about thought experiments. I suggest that these judgements constitute knowledge when we employ a competence (in the sense of Sosa [2007b]) in the application of the relevant concepts. I then appeal to some theories of concepts discussed in psychology to illustrate how such competence might be realized in our cognitive achitecture; The picture is not only consistent with the conclusions of the previous chapters, but in fact it supports them and it is supported by them in turn. It is part of the picture that the competence we use in applying concepts is shaped by our experiences, and it has to be distinguished sharply from the mere possession of the concepts.

In the last chapter, I turn to discuss a challenge to the use of thought experiments in contemporary analytic philosophy arising from the 'experimental philosophy' movement. I see that challenge as a new way of pressing an old worry, the worry that the sort of scenarios that philosophers consider in their thought experiments are often too far-fetched or bizarre to be the object of 'serious' reflection. I will distinguish different forms of the challenge. In its clearest form the argument proceeds from an empirical premise about the unreliability of non-philosophers' judgements about a certain class of thought experiments, and extends that result, inductively, to philosophers. In this chapter, I argue that there is no way of individuating the class of thought experiments under discussion which makes the argument both interesting and sound. Moreover, I will argue that the inductive step in the argument has not been justified; there are reasons to think that philosophers possess some sort of expertise in evaluating thought experiments which sets them apart from non-philosophers in relevant ways. Even though philosophers use the same reasoning capacity that non-philosophers use, we should not assume that they cannot become in some ways better at using it.

Chapter 1
Intuitions, Inclinations and their Epistemic Value

Introduction

Sometimes the debate on intuitions in philosophy is framed in terms of the question 'what are intuitions?'. In a certain sense, that is a basic question. But there is a sense in which the question is not basic, because it contains a presupposition, the presupposition that there are intuitions. We all have in mind paradigmatic cases of arguments in philosophy in which some judgement (usually, but not always, related to a hypothetical scenario) is thought to be an 'intuition', or considered 'intuitive'. If this is all is meant by 'there are intuitions', than it is a fairly harmless presupposition. But there is a different sense in which one could claim that there are intuitions, and it would then be a very substantial claim. The claim could be taken to be that there is, to use George Bealer's phrase, "a genuine kind of conscious episode"[2] which either constitutes or grounds these judgements. Furthermore, Bealer, like many others, believes intuitions play, and should play, a crucial role in our epistemic practices in general, and in philosophy particularly.

I want to consider some reasons that are often offered to defend these substantial claims, and show they are not cogent. There are of course other possible views, on which the claim that there are intuitions is still substantial, although it does not commit one to any mental kind over and above those already recognized; e.g. one might claim that an intuition is just a belief that has a particular source, or a particular kind of epistemic ground. I will not yet argue for or against any of these views in this chapter, but it will prove important to keep this possibility in mind at certain points. I will consider two broad kinds of arguments for the set of views under discussion. The first kind of argument relies on the fact that intuitions are needed to play some kind of epistemic role, so that a view on which they do not exist, or they do not play that role, ends up being self-defeating. The second kind of argument relies on

[2] Bealer (1992) p. 99

phenomenological considerations, and in particular on the role of seemings. I will grant that we can individuate a rather specific conscious state which we might call a seeming, but I will identify it with a felt inclination to believe. I will then respond to arguments to the effect that seemings must play some crucial epistemic role in philosophy and elsewhere.

1.1. The Coherence of Doing without Intuitions

The first kind of argument we will discuss goes back (at least) to Bealer's [1992]. Bealer [1992] is arguing against a radical, allegedly Quinean, form of empiricism, which he characterizes as the conjunction of three theses; the most relevant here are the first two:

> (i) A person's experience and/or observations comprise the person's *prima facie* evidence

And

> (ii) A theory is justified (acceptable, more reasonable than its competitors, legitimate, warranted) for a person if and only if it is, or belongs to, the simplest most comprehensive theory that explains all, or most, of the person's *prima facie* evidence (Bealer [1992] p.99)

For the purposes of the present discussion, it will suffice to say that the third assumption that Bealer is (reasonably) attributing to the radical empiricist is that we are justified in believing many things about the world; at least, and paradigmatically, some results of the natural sciences. I will take this to be entirely uncontroversial in this context. I will also grant here (ii), although I am going to raise some questions about the correct way to interpret some crucial terms in it.

Bealer thinks the problem lies in (i). Bealer presents three arguments for the conclusion that, assuming the other two parts of the Quinean picture, (i) has to be given up. I will sum them up here in a fairly rough way, since the details will turn out to be irrelevant. First, accepting (i) deprives one of necessary 'starting points' in theorizing, which are given by intuition; e.g., in order to treat observations as evidence, one would have to start with some beliefs about what counts as an observation and what does not. Bealer does not say anything general about what counts as a "starting point"; however, I think they could be thought of as justified beliefs which are based neither on perception nor on other beliefs. If this characterization is taken seriously, it is of course not uncontroversial that there are any starting points; still, it is quite obvious that there are *prima facie* candidates, such as simple

categorization judgments, and we can interpret the challenge as providing an account of what justifies those. A second argument relies on the claim that, by excluding intuition, (i) arbitrarily restricts our sources of evidence. In Bealer's analogy, it would be just like declaring our perceptual evidence to consist of auditory sensations only, excluding all the other sense modalities. So (i) is incompatible with the (in Bealer's terms) standard justificatory practice that is supposed to support our knowledge of the world, which includes appeal to intuition. A paradigmatic example of appeal to intuition is provided by thought experiments. The third argument is that the Quinean theory is directly epistemically self-defeating, in that (i) and (ii) do not come out justified if they are true, since experience cannot in principle give evidence in favour of normative claims such as (i) and (ii).

What is most interesting here is that the reasons Bealer gives against (i) are supposed to all depend on the fact that (i) excludes intuition from prima facie evidence. This is evident already in Bealer [1992], where he suggests the problems are avoided by substituting (i) with

"(i^i) A person's experience and intuitions comprise the person's *prima facie* evidence"[3]

Bealer [2000] makes clear that the arguments could be seen as attacking any view on which one's basic evidence does not include intuitions (and, *a fortiori*, any view on which intuitions do not exist). The general upshot is the following: "Whoever engages in reflective epistemic appraisal of their beliefs will end up in an epistemically self-defeating position unless they accept intuitions as evidence"[4]. I am mainly here concerned to argue against the latter conclusion, by showing ways of modifying (i) that do not appeal to intuitions; secondarily, I will also argue that in fact there is no incoherence in accepting (i) and (ii) as they stand.

In presenting the arguments, Bealer relies on a description of our standard justificatory practices that includes appeal to intuition. For example, in presenting the second argument, Bealer writes about a particular Gettier case:

> We find it intuitively obvious that there could be a situation like that described, and that in such a situation the person would not know that there is a sheep in the pasture despite having a justified true belief. This intuition - that there could be such a situation and in it the person

[3] Bealer [1992] p. 126. When he talks about the intuitions being evidence, Bealer means to allow the propositional content of this attitude to be part of the evidence (as opposed to the mental state itself) – I will do the same later when I will talk about, e.g., certain beliefs being part of the evidence.

[4] Bealer [2000] p. 7

> would not know – and other intuitions like it are our evidence that the traditional theory is mistaken.
>
> So according to our standard justificatory procedure *intuitions* count as *prima facie* evidence. (Bealer [1992] p. 100)

Now, in a certain sense what Bealer says is uncontroversial. There are some judgements about hypothetical cases which we find "intuitively" obvious, and they, or better their contents, are used as evidence. But I distinguished at the outset this sense of admitting that there are intuitions from the more interesting ones. If the more substantial sense is assumed, Bealer's description of the standard justificatory practice is question-begging, as I hope the discussion to follow will make clear.

The three arguments have a similar structure; there is a certain role that something has to play in our most comprehensive theory of the world (being a starting point, justifying basic epistemic principles, explaining our standard justificatory procedure). But perceptual experiences cannot play that role. Therefore we need intuitions.

The general form of argument, as one can see, presupposes that there is no further alternative; there is no further possible kind of *prima facie* evidence, or at least no kind playing the required role. Without this assumption, the argument against the radical form of empiricism cannot be turned into an argument for intuitions, as Bealer seems to think. But this assumption is controversial to say the least, and at any rate Bealer does not give us any reason to think it is true. This should be enough to suspend our judgement on the soundness the arguments, but I will go a small step further, with the aim to show that the assumption is almost certainly false. I will do this by showing that there are some alternative possibilities that Bealer, or anyone defending his argument, should consider. I will describe three views in particular of what might play roughly the required evidential role. None of the three views, as far as I can tell, has been ruled out. Not only did Bealer not rule them out, but the literature on intuitions did not either. I won't argue in favour of either of the three candidates, as that would be beyond the scope of this chapter. Chapter 4 will defend a version of one the views.

Before describing the three views which constitute a problem for Bealer's argument, I will put aside a different worry for Bealer's argument, based on a fourth alternative view. One could say that the role is played by beliefs marked by a link between understanding and (disposition to) assent. This view, although incompatible with radical empiricism of the sort

Quine defended, does not at all require positing a special sort of mental state[5]. However, I think this kind of view has been refuted – I will talk about it in chapter 2 at some length. Of course one may disagree with me on this assessment, but in this context, in putting this possibility aside, I am making it harder for me to resist Bealer's argument. So here I can put this kind of views aside, without any dialectical problem.

There are three more possible views which Bealer does not consider (I am not claiming the classification is exhaustive; there might be more). Firstly, one could reply, as I think Wright would (see e.g. Wright [2004a], [2004b]), that the role is played by beliefs which are not based on empirical evidence because they are not based on any evidence at all, although they are still in some sense warranted. One way of spelling out this idea is thinking of these beliefs, which we might call 'hinges', as pragmatically justified. We would not be in a position to undertake any cognitive project unless we had these beliefs. Therefore, they do not admit of reasons in their favour, being presuppositions of the activity of giving reasons. The view is at least prima facie plausible for our most general epistemological commitments. There is an issue about how widely it can be extended. In order to provide a rationale for a class of beliefs as wide as that of 'intuition', some kind of context-dependence of entitlement would be arguably required, letting different propositions count as hinges, or starting points, in different situations (this would be very close to the spirit of Wittgenstein's *On Certainty*). While the view was offered by Wright as a reply to certain forms of scepticism, it does not seem inappropriate to consider it here, as Bealer's arguments can be seen as suggesting that certain sceptical conclusions follow from abandoning intuition as a source of evidence. It is important here to note that Bealer, to make his overall argument cogent, should rule out not only each of the views I am discussing but also any plausible combination of them, which provides a different kind of solution to each of the three incoherence charges. Henceforth, I will put the idea of combining the different views aside, going along with the fiction that a single kind of state or entity has to play all the roles Bealer thinks intuition plays; this of course will make my task harder, and it will also simplify the discussion.

A second kind of view that provides something playing the required evidential role, without appealing to intuitions in Bealer's sense, is an (epistemologically) externalist view of the relevant judgements. Such a view has been developed e.g. by Sosa ([2007a], [2007b]) and Williamson [2007]. According to this kind of view, beliefs in simple categorization

[5] In a sense, Bealer still endorses the view, in requiring that "rational" intuition be based on understanding. But that part of his view suffers from the same defects of similar views, that will be addressed in chapter 2.

judgements, which Bealer thinks need a previous intuition to be supported, are justified insofar as they derive from a reliable source, i.e. they are formed by a subject through the exercise of competence such that the subject could not have easily formed a false belief. In Williamson's view, there is a vast class of propositions that we can know, to use his expression, from the armchair; roughly this means that we can know them without perceptual or empirical evidence playing any direct evidential role; the possibility is left open that experience plays a role in shaping our concepts and our ability to use them which goes beyond what has been traditionally characterized as an enabling role. A paradigmatic example of this phenomenon is, according to Williamson, knowledge of counterfactual judgments. Williamson applies this analysis in particular to one of the examples which Bealer used in the second argument, the Gettier cases, in which the proposition playing the key evidential role would then be, roughly, the one expressed by "If someone were in the situation described, she would have a justified true belief but that she would not have knowledge".

What is important for us about the view are the two following features: the sort of cognitive capacities employed in gaining this kind of knowledge is, according to the view, absolutely ordinary; not only is no special faculty involved, but the resulting knowledge lacks any epistemologically interesting distinctive mark. Secondly, this kind of judgements is not usually (or at least, in some central cases is not) inferential in any interesting sense. We can make the difference with Bealer's view more vivid by comparing (i^{i}) ("A person's experience and intuitions comprise the person's *prima facie* evidence") with

(i^{ii}) A person's perceptual experiences and reliably formed beliefs comprise the person's *prima facie* evidence

To introduce the third view which would permit avoiding Bealer's arguments, consider the following further principle which could replace (i) without yielding contradiction.

(i^{iii}) A person's perceptual experiences and beliefs comprise the person's *prima facie* evidence

To count as evidence, arguably, a belief should be justified; therefore, one might object, the truth of (i^{iii}) requires that all beliefs are *prima facie* justified. The view that all beliefs are *prima facie* justified is Epistemic Conservatism (henceforth EC). This view was defended

very clearly in Harman [1986], [2001] and [2003][6]. Harman also claims that the account he is offering is "one [he] think[s] it is clear that Quine accepts, apart from terminology"[7]. I agree on the latter point. I take that to mean that Quine thought the existing beliefs played a role in spelling out (ii) above. To decide what the simplest explanation of the evidence is, we need to consider, among other criteria, which is the most conservative; other things being equal, an explanation preserving our current beliefs is better than an explanation requiring us to change them (so what explanation of a set of experiences is best will be relative to what background beliefs we hold). This is particularly interesting for us, since it allows us to formulate the reply to Bealer's arguments in two ways, which are arguably equivalent. The first is advocating (i^{iii}) as a substitute for (i); the second is leaving (i) and (ii) as they are and reading (ii) as I just suggested, so that other things being equal, an explanation that includes existing beliefs is simpler than one which does not[8]. In any case, Bealer's incoherence charge is directed to a straw man, insofar as it ignores the role of conservativeness in assessing a theory. This is particularly clear in thinking about the "starting point" argument. Our starting point, according to Quine, is where we are[9], i.e. the totality of our current beliefs[10]. If we take

[6] Harman dubs his positions in various other ways, while, as far as I can see, the position has remained substantially unchanged; in his [1986] he classifies it as a form of coherentism, while in [2001] and [2003] he counts it as a form of foundationalism ('generalized' foundationalism).

Bealer mentions coherentism in the last part of footnote 3 in his [1992], just to say that he is not arguing in that paper against it. If one wants to count the view encompassing (i^{iii}), (ii) and EC as coherentism, then I believe Quine was a coherentist, and therefore the mentioned footnote is inconsistent with the stated purpose of Bealer [1992]. If one does not want to count the view as coherentism (and I believe this is the better terminological choice), then Bealer's arguments fail by not considering at all this option.

[7] Harman [2001] p. 659. See also the references in fn. 9

[8] Or, to make things clearer, we could modify (ii) by substituting "best" for "simplest".

[9] See Quine [1960] (pp. 3-5), Quine [1970], (pp. 7, 86, 100), Quine and Ullian [1970] (ch. 6), Quine [1995] (p. 49).

[10] A further terminological variant of the view has been defended by two more influential authors who are quite explicitly inspired by Quine from a methodological point of view. According to David Lewis and Peter van Inwagen, what we call "intuitions" are usually beliefs (or inclinations to believe). For example Lewis [1983], p. x, writes: "Our 'intuitions' are simply opinions; our philosophical theories are the same. Some are commonsensical, some are sophisticated; some are particular, some general; some are more firmly held, some less. But they are all opinions, and a reasonable goal for a philosopher is to bring them into equilibrium". Van Inwagen [1997], p. 309**: "**Our intuitions simply are our beliefs – or perhaps, in some cases, the tendencies that make certain belief attractive to us, that "move" us in the direction of accepting certain propositions without taking us all the way to acceptance (philosophers call their philosophical beliefs intuitions because 'intuition'

this point into account, the coherence of the Quinean view is only threatened by the following version of the third argument. According to a certain interpretation of Quine's view, which Bealer seems at some point to take for granted, the whole epistemic vocabulary would have to be eliminated from our language. But Quine, according to Bealer, holds (i) and (ii), which contain epistemic vocabulary. Now, the very fact that on this interpretation Quine's view becomes *blatantly* inconsistent, should constitute exegetical evidence against it. So one would expect Bealer to provide at least some textual grounds in its favour. But Bealer is open about the fact that he is not doing that (Bealer [1992] fn. 2). Be that as it may, I am, as I said above, mainly concerned with the coherence of views which do not accept intuitions, and secondarily with the coherence of (i) and (ii), and not with the coherence of Quine's actual views.

Bonjour [2001] replies[11] to Harman claiming that the view implies we can be justified in believing whatever we want. If that were so, the view would indeed have to be discarded as a serious alternative. Of course the problem was anticipated by Harman; I quote from Harman [2001]:

"Objection: 'In this view, as soon as one randomly comes to believe *P*, one is automatically justified in believing *P*.' Reply: If one believes not only *P* but also that one randomly came to believe *P*, the two beliefs are in tension and one has a reason to abandon at least one of them."[12]

The undesirable consequence still follows, according to Bonjour, provided we are able not only to form the belief, but also to change whatever "logical or methodological views"[13] make the new belief be in tension with the others we have. An immediate reply could be that we are not in fact able to change our beliefs at will, and creatures that were able to do so would be so different from us as to make irrelevant what would count as justified for them. However, I think the objection would be better met by noting that a defender of EC is not committed to saying that what counts as a tension between our beliefs depends on what we think is a tension between our beliefs. Perhaps it is part of Harman's theory of meaning, and perhaps it is even a possible interpretation of Quine, that there is a tension only if our current epistemic practices as a whole determine a tension. However, nothing in EC or (i) and (ii) commits one

---

sound more authoritative than 'belief')". On this view, we could drop (i) and accept (B), but this clearly would not amount to admitting the existence, and much less the epistemic significance, of intuitions in Bealer's sense.

[11] Harman [2001] puts forward the view as a response to Bonjour [1998]; the view defended in Bonjour [1998] has a clear connection with Bealer's view, and with the sort of argument in its favour we are considering here.

[12] Harman [2001] p. 658

[13] Bonjour [2001] p. 692

to such views about meaning. Plausibly, there are objective logical and evidential relations between the contents of our beliefs, and our practice should conform to them. Again, I was not concerned here with defending Quine's or Harman's view, I only aim to establish that there are plausible alternatives to Bealer's proposed modification of (i). It is worth noting that EC is plausible only if the notion of justification employed is interpreted in an internalist sense, or at least that is how Harman intends the notion to be interpreted (see Harman [2003]). This does not make it less relevant for Bealer's arguments, which are concerned with rational coherence. If internalist notions of justification have any application at all, it is precisely in dealing with matters of rational coherence.

Of course, I have not provided here reasons to think EC is true, as I haven't done for the other two views sketched above. My point is that for the sort of arguments Bealer gives to work one would have to exclude all views in which something other than intuition plays the required evidential role and, importantly, any combination of them. Moreover, on a charitable interpretation of (ii), Bealer's argument does not even show that the radical empiricist view, in the formulation chosen by Bealer himself, is incoherent. What this suggests is that not only does Bealer rely on an assumption he did not defend, but also that providing such a defence is going to be a daunting task, and one we can reasonably assume cannot be carried out.

1.2. Beliefs, Intuitions, and Dispositions

Let's now move on to the other style of argument, the one based on phenomenological considerations. Here is one formulation of the rationale Bealer repeatedly (Bealer [1992], [1996], [1998], [2000], [2002], [2004]) gave to think of intuitions as a distinct mental state:

> By intuitions we mean *seemings*: For you to have an intuition that p is just for it to *seem* to you that p. Here 'seems' is understood, not in its use as a cautionary or "hedging" term, but *in its use as a term for a genuine kind of conscious episode*. [My italics] For example, when you first consider one of de Morgan's laws, often it neither seems true nor seems false; after a moment's reflection, however, something happens: it now just seems true. This kind of seeming is *intellectual*, not experiential—sensory, introspective, imaginative.
>
> Intuition is different from belief: you can believe things that you do not intuit (e.g., that Paris is in France), and you can intuit things that you do not believe (e.g., the axioms of naive set theory). The experiential parallel is that you can believe things that do not appear (seem sensorily) to be so, and things can seem sensorily in ways you do not believe them to be (as with the Müller-Lyer arrows). (...) Now, since intuition is analogously different from other

> psychological attitudes (judging, guessing, imagining, etc.) and from common sense, I believe there is no choice but to accept that intuition is a *sui generis* propositional attitude. (Bealer [2004], pp. 12-13)

Is there a use of 'seems' "as a term for a genuine kind of conscious episode"? and if there is such a use, is there a genuine kind of conscious episode to be picked up? Bealer seems[14] to take for granted a positive answer to both questions, rather than argue for it. If one does take that much for granted, one can then define intuition in terms of seemings[15]. But the grounds for the positive answer are shaky at best. I will attempt to provide in this section an account of what it takes for something to seem to be the case which does not posit any *sui generis* mental state.

Given that I cannot escape discussing the appeal to introspection, I have to testify that I do not remember any aspect of my conscious life common to all episodes of seeing, e.g., the correctness of a theorem, except perhaps a certain sense of satisfaction in having grasped something I did not grasp before; much less can I think of something in common between the phenomenology of that grasping and, say, the phenomenology of the judgement that a subject in a Gettier case does not know.

Sometimes, I grant, it is true that it seems to me that something is the case; does that force one to admit a genuine kind of conscious episode which makes it true, the 'seeming' that I enjoy? It is interesting that such a noun has no common use in English. However, we certainly can stipulate that a seeming is a mental state which occurs in a subject just in case something seems true to the subject. But we are still quite far from the need of positing a "genuine kind" in any interesting sense of that expression. English is a rich language, and there are many ways of expressing propositional attitudes with, as we say, a world-to-mind direction of fit: sometimes it seems to me that *p*, sometimes I believe that *p*, I believe that *p* to some extent or other, I am confident that *p*, I have (a certain degree of) confidence that or in

---

[14] This "seems" has a hedging function.

[15] There is a complication here, for Bealer, which I am not discussing. He does not want all seemings to ground intuitions in the relevant sense, but only "rational" ones. The way he differentiates rational seemings is by the fact that they present their content as necessary. But this will not do, as noted by Goldman [2002], because one can have intuitions without having the corresponding necessity intuition, and indeed without having modal concepts at all. See chapter 3, section 3.3 for further discussion.

.

p; moreover, I often think that *p*, I am sure or certain that *p*, I have an opinion or a conviction that *p*, I take for granted that *p*, I presuppose that *p*, I have a feeling that *p*, a hunch that *p*, I guess that *p*, I surmise or suppose that *p*, I suspect that *p*, I conjecture that *p*; sometimes I find it plausible or verisimilar that *p*, or it looks to me as if *p*, and we could go on. All of the foregoing can be used to express some positive psychological attitude toward the truth of a proposition. To posit a genuine kind of conscious episode for each term would certainly be a crazy way of doing psychology. Surely it would be equally bad practice to assume all the foregoing must be analyzable in terms of one fundamental kind; but it should be clear that, other things being equal, an account which posits fewer fundamental states should be preferred, on grounds of simplicity. Two obvious candidates for that role are belief and credence; and an equally obvious candidate is a combination of both. Good old all-or-nothing belief is the one option I would go for, but it is not my purpose here to defend that view in general. What I want to stress is that, other things being equal, positing a new "genuine kind" is a cost.

Still, Bealer's examples show, I think, that there is no simple equivalence between seemings and beliefs. Unless one is an eliminativist about seemings, someone who thinks nothing ever seems to be the case, an analysis will have to be provided. This will be my aim in the rest of this section. But this should be qualified in various ways. I am trying to clarify a particular use of 'seem', one in which it expresses a rather specific propositional attitude, and I am neutral on the relations that there are between this and other kinds of use of the word, like the hedging use, mentioned by Bealer, or the objectual use (as in 'you seem tired') and the purely perceptual use (as in 'in the Muller-Lyer illusion case, it seems that the two lines are of different length'). It might be that 'seem' is ambiguous in English (in that case, I would surmise that the perceptual use is the primitive and literal one, and the others are figurative extensions); or it might be that 'seem' covers unambiguously all the uses I mentioned, but what we focus on here is *a specific kind* of seeming.

Let us discuss first the case of seemings in the absence of belief. The example typically offered is that of a proposition which seems true although we know it to be false; as a case of this kind we are often offered the naïve axiom of comprehension. Even after you have learned that it leads to contradiction, you can continue to find it intuitively attractive, so to speak; it still seems to you, roughly, that to any property will correspond a set of things having it. But we might also have different cases. Weatherson [2003] gives an interesting example of a case in which you might have the intuition that *p* without either believing it or believing its negation. Consider the statement 'in a fair lottery with 1,000,000 tickets, I do not

know that ticket 1 is not going to win'. You might have the intuition that this is true, and at the same time worry that if it is then we might have to accept scepticism, and therefore, while you think about the matter, you still have the intuition but you suspend your judgement.

Williamson [2007] suggests that in all these cases, although there is no belief, there is a felt inclination to believe[16]. I will defend the view that seemings, in the sense we are interested in, just are felt inclinations to believe.

The defender of seemings as a *sui generis* mental state is here likely to object that some explanation is required about what 'felt inclination' means. A possible reply is that the expression 'felt inclination' is not used in any technical sense. Ultimately, I think that is correct, and the natural meaning of the expression will do the job. Nevertheless, I feel a request for some clarification of what that meaning amounts to is legitimate, and I will try to provide it. Therefore, I will say something in turn on what an inclination is, and on what it takes for an inclination to be felt (as far as I can see, Williamson could agree with everything I will say, but he certainly leaves the matter open).

A first rough and ready characterization of what I mean is the following: An inclination, for our purposes, is just a disposition. The disposition to believe, in order to prompt the judgement that its object seems to be the case, should be felt; this means, roughly, that you should be aware of the disposition, not just in any way, but rather in the particular way you are aware of some of your mental states. Let's consider difficulties for the two parts of this characterization in turn.

There should be no doubt that giving a full account of what a disposition generally consists in is a difficult and up to now unfulfilled (or not completely fulfilled) task. There should be equally no doubts that there are dispositions, and human beings in particular have some. It would not therefore be reasonable for an objector to ask for a complete theory of dispositions, and much less to ask for a reason to believe there are dispositions.

There are still two complaints that I see as worthy of consideration. The first is that one might want to know something about this particular kind of disposition, e.g., which are its manifestation conditions. Secondly, one might complain that there is no disposition at all to believe in some cases of intuition without belief, such as the Naïve Comprehension Axiom case. Given that I know it is false, we would not normally say I have any disposition to

[16] Sosa [2007a] makes a similar suggestion about intuitions, claiming that an intuition is a conscious state of attraction to assent. Not all such states however are counted as intuitions in Sosa's theory, but only those meeting a certain number of additional constraints. The account offered here differs first because its object are not intuitions but rather seemings, and secondly because there are no additional conditions.

believe it. Of course, the objection would go, there are trivial ways to specify conditions under which I would believe it to be true, even if I actually know it to be false. For example, if the evidence was that it is true, and I had no reason to doubt the evidence, I would believe it; but this is true of any proposition, or nearly any.

The latter objection however is not very worrying. First, as Williamson observes, an inclination can be present even in situations in which there is no risk that it will be manifested: "I can feel such an inclination [to believe Naïve Comprehension] even if it is quite stably overridden, and I am not in the least danger of giving way to temptation (just as one can feel the inclination to kick someone without being in the least danger of giving way)"[17].

It is also important to keep in mind that the manifestation conditions for the subject's disposition to believe need not be the same in every case. Talk of dispositions is somehow context-dependent. Consider the standard example of 'fragile'. What it takes for a glass to be fragile is not the same as what it takes for my ankle to be fragile. It's not just that my ankle will resist a stronger pressure. After all, a force much weaker than the one produced by the weight my ankle normally bears could suffice to break it, if applied in the right sort of way, as a twist. Just like the manifestation conditions for fragility can vary, so can those for inclination to believe, and even inclination to believe a specific proposition. In other words, 'inclination' is a context-sensitive term. Keeping this in mind, we can easily think of contexts in which we would say we are in fact inclined to believe the Naïve Comprehension Axiom. Surely I am more inclined to believe that than to believe that 2+2=5.

Moreover, the defender of seemings is not here in a position to coherently deny the presence of the disposition. For the seeming is supposed to be a primitive form of evidence; and evidence surely does support belief, if one is rational. So there should be, according to their view, a manifestation condition avoiding the triviality problem, in some contexts at least, namely the lack of any independent (independent of the disposition itself) evidence in favour or against the proposition, and normal rational behaviour.

We could start looking for counterexamples to this or other proposals about manifestation conditions, in the form for example of true counterfactuals of the form 'If I were to lack any independent evidence in favour or against *p*, and to be normally rational, I would believe *p*' in situations in which we don't want to say there is a disposition to believe. But I would like to urge the reader to consider again that we are here not presupposing any theory of dispositions, and in particular no simple connection between dispositions and

[17] Williamson [2007] p. 217

counterfactuals. Most theorists think that conditional analyses of dispositions are riddled with counterexamples. These take the form of different kinds of possible situations in which an object has a disposition but if the (normally appropriate) manifestation conditions were to obtain the object would not manifest its disposition (see e.g. Fara [2006] and references there). One can react to this problem in various ways. Fara [2005] propounds an analysis of dispositions in terms of generics; something has a disposition to Φ if and only if (generically) it Φs when the manifestation conditions obtain. An analysis in terms of a generic trades informativeness for adequacy; it gets rid of all (or most) counterexamples, but it risks saying very little. It is worth noting that many theorists recently defending counterfactual analyses often accept a similar trade-off, for example by offering a conditional which is weakened by the inclusion of a *ceteris paribus* clause in the consequent (see e.g. Steinberg [2010]). In a different way, this point also applies to the theory defended by Manley and Wasserman [2008], on which

"Prop *N* is disposed to *M* when *C* iff *N* would *M* in some suitable proportion of *C*-cases"[18]

Manley and Wasserman's theory adds to the simple counterfactual analysis a sort of context-sensitive ceteris paribus clause. I will not here endorse any particular theory of dispositions, but, by way of exposition, it will be useful to adopt Manley and Wasserman's theory. Applied to our proposed manifestation condition, this gives us that you have a disposition to believe the axiom of comprehension iff you would believe the axiom of comprehension in a suitable proportion of cases in which you lack any independent evidence for or against it and you were being normally rational. This avoids most obvious problems. Also, it elegantly explains at least one dimension on which the strength of your seeming could vary, while only using an all-or-nothing notion of belief. Of course, any good theory of dispositions should account for the fact that dispositions vary in strength, and therefore lend an explanation of the corresponding gradability of seemings.

Notwithstanding all that, I predict some readers will still want to go and look for counterexamples to any particular manifestation conditions one might propose, and I am not ruling out that they might succeed. If that happens, however, before giving up on the idea of understanding seemings on the line of felt inclinations, it will have to be excluded that the

[18] Manley and Wasserman [2008] p. 76

fault is likely to lie rather in the proposed manifestation condition or in Manley and Wasserman's account.

So far, I have addressing counterexamples to the necessity of a felt inclination to believe for a seeming. Now we will look at some alleged counterexamples to the sufficiency direction. It might be objected, to start with, that not all dispositions to believe will do, because some of them lack the sort of phenomenological character required to count as evidence. I take that to be a worry worthy of serious consideration. I will consider three different versions of this worry. The first two will be addressed in this section; the third, which I take to be the most interesting one, will be addressed in the next section.

Suppose I have a disposition to believe *p* that is completely unknown to me, and I deny that it seems to me that *p*. In such a case, my denial can be correct. This problem can be easily dismissed by requiring that I am aware of my disposition. However, this is arguably not going to be enough. Here is where the first form of the worry I mentioned arises: suppose I claim that I do not find the axiom of comprehension intuitive at all. I was exposed to it before being aware of the contradiction you can derive from it, and did not believe it. However, a neuroscientist, whom I have very good reasons to trust, comes to me with a device of her creation which can read dispositions to believe off people's brains. She goes on to tell me I do have a disposition to believe the axiom now. Suppose also this is true, and the neuroscientist's device is perfectly reliable. I now know, plausibly, I have a disposition to believe, and yet it does not seem to me that the axiom is true.

This just goes to show that not every form of awareness of my disposition is sufficient for me to have a seeming. In Williamson's term, the disposition must be "felt"; I take this to mean that I must have some sort of introspective awareness of my disposition. In common talk, something is rarely called an 'inclination' unless it has this felt character, so I will here use 'disposition', to avoid making the specification redundant. We do have this sort of awareness for other kinds of dispositions as well; dispositions to act in a certain way, or to get into a certain mental state. You might feel that you have a disposition to get angry at a certain person, or you might feel that you have a disposition to go for a walk, or a disposition to start crying very soon, or a disposition to fall asleep, and so on. Of course, this does not imply that you have (in each case) the corresponding desire. In these cases as well, knowing by testimony you have the disposition would result in a different mental state. Of course, it is a good question, for philosophers and psychologists alike, how we achieve introspective awareness of these things, and I do not have a full answer to that. Maybe the introspective awareness is not of the disposition directly, but rather of a certain phenomenological state we

associate with the manifestation of the disposition; for example, if I experience a certain feeling I had in the past before falling asleep, I might come to know through it that I have a disposition to fall asleep right now. Be that as it may, there is nothing particularly surprising in our being aware of dispositions to believe. It would be surprising, indeed, if inclinations to believe were different.

We also have to notice, as this will be relevant later, that the fact that you have a disposition to believe a proposition will sometimes count as evidence in favour of the proposition. Suppose you are undergoing some sort of psychological experiment, and you are told that, if you had not taken the pill you have taken, you would find obvious a certain claim which you feel at present no disposition to believe. Suppose the claim in question is some simple mathematical statement, or some claim of synonymy between two terms you are familiar with. It would be rational for you to conclude the claim is probably true. This is because in this case the claim belongs to a class of contents about which you justifiably think your dispositions to believe are usually a reliable guide to truth. If the claim is not in any such class, then learning that one has a disposition to believe by testimony is not (good) evidence for its truth, but neither would a seeming be. Suppose it seems to you that the number of cats in New York on the 14th of March 1877 was even. That surely wouldn't count as good evidence. Similarly, learning that you have a disposition to believe it wouldn't provide evidence to believe it. So, although seemings cannot be identified with dispositions to believe, since the introspective awareness is needed, it is not clear that the nature of the awareness plays any important evidential role. We will come back to this in section 1.3.

There are two more challenges to the sufficiency of felt inclination to believe for seeming which I wish to consider. A very interesting kind of putative counterexample can be extracted from Tolhurst [1998]. Tolhurst considers a case in which the disposition is felt, but it is unrelated to the truth of the proposition, and it therefore allegedly fails to constitute a seeming. Here is the case, and Tolhurst's conclusion about it:

> Case III: Fred's son, Fred Jr., has been charged with [a horrible crime]. The evidence is overwhelming. Even so, Fred feels an irresistible inclination to believe Fred Jr. is innocent; he just cannot believe that his son could have done such a thing even though Junior's past behaviour provides clear evidence that he could.
>
> Fred's overwhelming desire that Fred Jr. be innocent inclines him to believe that Fred Jr. is innocent. But this desire is not a seeming even though it inclines Fred to believe its content (Tolhurst [1998] p. 297)

I agree of course that the desire is not a seeming. My view however entails that a seeming (caused by the desire) is present in this case, since there is an inclination to believe, and presumably the subject is introspectively aware of it. Tolhurst thinks that, while every seeming is accompanied by a disposition to believe (so he is committed to rejecting the counterexamples to the identification of seemings and felt inclination we discussed before, in which allegedly there is a seeming but no disposition to believe), not every inclination to believe, even if it is felt, constitutes a seeming, and the case described is supposed to provide an example of that.

The first thing to observe about the case is that it is under-described in one important respect; we do not know whether Fred appreciates the strength of the evidence. Take the case in which Fred does not appreciate the strength of the evidence. Then he is misled by his desire; it seems to him that his son his innocent, even though it wouldn't seem so to an impartial observer. There is no difficulty here, because seemings in general do not have to track the subject's total evidence. On the other hand, suppose Fred fully realizes the evidence against Junior is overwhelming, but he believes he is innocent (or suspends judgement) because of his inclination, although he is also aware that this inclination is purely irrational. This is a tricky case, because some theorists would deny it is possible. Beliefs are not (or at least, mostly not) under our voluntary control; I cannot right now decide to believe that the number of stars is even, or that I am in Paris, given that I know I do not have evidence for these claims. According to Williams [1973] this is not a contingent fact about our psychology, but rather something following from the fact that beliefs aim at truth. Therefore, it is not possible to consciously believe against your evidence, and also, I suppose, to consciously suspend belief against evidence. Without taking this extreme view, I think considerations similar to the ones offered in its support are sufficient to explain the oddity involved in this case. If you realize that the source of your disposition to believe something is utterly independent of the truth of its object (note how easily you can be misled on this; often people believe their "feelings" are a reliable guide to truth, even when they should not), this surely not only renders the disposition evidentially irrelevant, but it also tends to undermine the existence of the disposition. This connects with Tolhurst's contention that the mark of genuine seemings is "the feel of truth, the feel of a state whose content reveals how things really are"[19]. I agree with Tolhurst that felt veridicality is the mark of seemings; I do not agree with him in taking it to be a sort of primitive concept. Given that to believe something is to regard it as true, felt veridicality is exactly what I would expect to mark the content of a felt

[19] Tolhurst [1998] pp. 298-99

inclination to believe. In the described case, as we take away the felt veridicality, we tend to cancel out the seeming, but also the inclination to believe (although of course it has not tendency to cancel out the subject's desire that the content of the belief is true). To sum up, on some specifications of the case it seems to Fred that his son is innocent, and on some other specification he does not have an inclination to believe it. Either way, there is no counterexample.

A final kind of putative counterexample to the sufficiency of felt inclination to believe for seeming might be extracted from the sort of cases which constitute a counter-example to the sufficiency of belief *simpliciter* for seeming, that is, beliefs whose content do not seem to be the case. Given what I argued so far, this is not an objection to my view unless we also think that every belief must be accompanied by a felt inclination to believe. This is clearly not so. One can make a judgment about some subject matter without having any introspective awareness of having a disposition to make the judgement; one could lack altogether the capacity to have introspective awareness of one's dispositions to believe, and still have beliefs. In fact, I count as an advantage of my view that it does not imply that every belief is accompanied by a seeming.

To illustrate, consider the following question: Does it seem to you that Paris is in France? Bealer suggests that it does not, and I agree. Similarly, I believe it would (under normal circumstances) be wrong to say that you *feel* an inclination to believe that Paris is in France. In general, it will not usually be appropriate to say "it seems to me that *p*", when you believe *p*. In some of these cases, the inappropriateness will be easily explained along Gricean lines. Just as it is often inappropriate to say 'I believe that Paris is in France', when you could say 'Paris is in France', it will often be inappropriate to say you have an inclination to believe *p* when you believe *p*, because the latter claim will be more informative. Similarly, one could try to explain pragmatically how the negation of the weaker claim is sometimes appropriate, like when one says 'I do not believe that, I know it'. But this strategy does not apply equally well to all cases. For example, the law of non-contradiction is something I believe, and it also seems true to me; by contrast, "some reptiles are not carnivorous" is also something I believe, but it does not seem true (or false). I thought about it, and I remembered that turtles are reptiles and they are not carnivorous, so I have a firm belief. But I do not have an independent inclination to believe it. Similarly, e.g., for the proposition that Strasbourg is in France, which often seems false to me, before I remember the relevant geographical and historical facts.

If the foregoing is correct, should we say that intuitions are just seemings? I believe that the use of the term 'intuition' in the philosophical community is not clear enough to

ground an answer[20]. Because of that, I believe the right question to ask is: should we use the term 'intuition' to refer to felt inclinations to believe? I favour a negative answer to the latter question; I do not see any gain in restricting the use of the word 'intuition' to felt inclinations to believe. A first reason is that philosophers often talk about people's intuitions in a way that does not require them to be thinking about the contents of the intuition. Surely when one is engaged in a philosophical project one might have to try to respect many intuitions at the same time; but one cannot feel many inclinations at the same time[21]. A second reason is that, insofar as beliefs without seemings are possible, there are occurrent beliefs which some philosophers want to count as intuitions although they are not accompanied by any occurrent seeming; e.g., as noted in Williamson [2007] (ch. 7), the judgement that I weigh more than my brain, or the judgement that there are mountains are counted as intuitions. It is also worth noting, in the context of this work, that seemings do not have any special relation to thought experiments. Certainly, not all seemings have as their content hypothetical judgments. Moreover, I do not think that all hypothetical judgments are accompanied by a corresponding seeming. If we wish to use the term 'intuition' in philosophy so that it is tied by definition to thought experiments, it is therefore not convenient to define 'intuitions' as just seemings. But, whether or not the term 'intuition' properly applies to seemings, the most important question is whether seemings play the crucial evidential role that, e.g., Bealer claims they play, in accounting for our justified judgements in at least some central cases. I will discuss this issue in the next section. I think seemings do not constitute either a *sui generis* or a particularly central kind of evidence in Philosophy. In the next section, I will defend the latter claim.

[20] To argue in favour of this claim would require a careful examination of the use of the term 'intuition', an related ones, in a vast amount of contemoprary philosophical texts; this would be very interesting, but it is not the aim of the present work. This work is about philosophical methodology, and not about the use of the term 'intuition' in philosophy, or even in the literature on philosophical methodology. I refer the reader to Cappelen [forthcoming] for a good start on a careful examination of the use of the term in philosophy, one which supports my claim above.

[21] This point could be accommodated pointing to a dispositional sense of "it seems that", corresponding to the dispositional sense of belief. Maybe we want to say of someone who is asleep that it still seems to her, e.g., that 2+2=4. I find that quite unnatural. In the theory that would correspond to ascribing a disposition to have a felt disposition. However, the consideration that in general being in state does not entail having a disposition to be in that state would still count against the general thesis that every belief is accompanied by a seeming in my sense.

1.3. Which Evidential Role for Seemings?

I think seemings do not constitute either a *sui generis* or a particularly central kind of evidence in Philosophy. In this section, I will defend the latter claim. It is quite important though to clarify that I take the burden of proof to be on my opponent at this point. It is not enough, to claim that seemings have some special epistemic role in philosophy, to point at the use of thought experiments. Given that seemings are to be distinguished from beliefs (although I argued that they are to be understood as inclinations to believe), surely it is possible for thinking subjects to use thought experiments, hypothetical scenarios, in various ways, even if they had no seemings. They could just form beliefs about the hypothetical scenarios. Actual philosophers certainly have beliefs about some hypothetical scenarios. Hence, the practice of appealing to thought experiments by itself, even assuming it is in good epistemic standing, does not require seemings to have a central epistemic role.

I will consider in this section two arguments that aim to establish that seemings have an evidential role which does not depend on anything else; in particular, it does not depend on other justified beliefs the subject has. On this view, seemings are the (or at least one) ultimate source of justification. A seeming with content *P* provides prima facie justification for *P* independently of any other justified belief the subject has. This kind of view is sometimes called *dogmatism* about seemings, or *phenomenal dogmatism* (Chudnoff [2011]). Dogmatism can be defended with respect to different epistemic sources; in particular, it is often proposed for perception. Dogmatism about any epistemic source faces some serious objections; it has been argued that it is not compatible with well-established probabilistic models of evidential relations, and it commits one to the acceptability of arguments which are intuitively circular (White [2006], Wright [2007]). I will not discuss those objections here. I do not think we need to, since the motivations for dogmatism about seemings are too weak to make it a serious contender.[22]

The first argument I will discuss, proposed in Huemer [2007], is similar to Bealer's arguments which were rebutted in section 1, in that it is a self-defeat argument. However, it also different in being concerned exclusively with seemings, rather than intuitions, and in

[22] Chudnoff [2011] does not defend dogmatism about seemings; rather he assumes that it is true, and he then tries to give a metaphysical picture which explains its truth. I believe he manages to provide such a metaphysical picture, although whether the picture is attractive is a different matter. In fact, one might think that the metaphysical picture, although not inconsistent, is so implausible and convoluted to constitute, if it is the only one which can explain the truth of the form of Dogmatism Chudnoff favors, a strong argument against such view.

relying, as we will see, on some additional empirical assumptions. Huemer endorses what he calls *Phenomenal Conservatism*, which he states as follows:

"PC: If it seems to *S* that *p*, then, in the absence of defeaters, *S* thereby has at least some degree of justification for believing that *p*" [23]

The expression 'it seems' is elucidated as follows "I take expressions of the form 'It seems to *S* that *p*' or 'it appears to *S* that *p*' to describe a kind of propositional attitude, different from belief, of which sensory experience, apparent memory, intuition, and apparent introspective awareness are species. This type of mental state can be termed an 'appearance'."[24] However, no argument is offered to prove that this is a unified kind in some interesting sense, and no further clarification is given to the term 'intuition' which helps distinguish intuition from other kinds of seemings. Huemer does argue instead against the view that some kinds of appearances confer justification, but not all. I will leave aside both issues for the sake of the argument. Let us look at Huemer's argument in favor of PC. Huemer sums it up as follows:

> If, that is, appearances do not confer at least some defeasible justification on propositions that are their contents, then since our beliefs are generally based on what seems to us to be the case (the reason we believe what we do is that it appears true to us; our method of forming beliefs is to believe what seems true to us), our beliefs are generally unjustified. (Huemer [2007] p. 41)

The crucial assumption here is that "our beliefs are generally based on what seems to us to be the case". As noted in DePaul [2009], since the basing relation (as Huemer – standardly - conceives it) requires a causal link between the belief and what it is based on, this a psychological claim which one would expect empirical evidence for, but no such evidence is offered by Huemer. Presumably, Huemer thinks there are phenomenological reasons to accept it, and he thinks they are so strong to make the claim obvious. But, in fact, I believe the claim is quite obviously false. The assumption actually comprises two distinct theses. The first is that each belief is accompanied by a seeming (presumably, one with the same content); the second is that the belief is based on the seeming. I will argue against the two components in turn.

Even granting to Huemer that perception, memory, introspection and intuition all present us with seemings, and that one can usefully generalize about this kind, there are two

[23] Huemer [2007] p. 30

[24] Huemer [2007] p. 30

kinds of belief that constitute clear counterexamples to Huemer's claim that every belief is accompanied by a seeming. First, there are inferentially justified beliefs, like the belief in Pythagoras' theorem, or my belief, mentioned above, that not all reptiles are carnivorous. The theorem itself does not seem to be true (in fact it is quite surprising that it is, so if anything the theorem itself seems untrue), and neither does the other belief. Huemer needs to argue that these beliefs are based on their seeming to follow from other things that seem to be true, and then reformulate the argument accordingly. A second kind of counterexample stems from beliefs that are stored in semantic memory (not memory of events of our own life, but rather memory of facts about the world). I am not aware of any seeming, for example, for the belief that turtles are reptiles, or that Paris is in France, just like for countless others; I am only aware of the belief. I am relying on introspection here, but Huemer can hardly complain about that, since his crucial empirical assumption is only based on introspection. Moreover, as we noted above, my introspective judgement is not anomalous or isolated. In fact, I am not aware of a single theorist besides Huemer defending the view that every belief is accompanied by a seeming. So Huemer is in a terrible dialectical spot; if introspection is reliable way to form beliefs about this matter, then we should go with the majority view, and conclude that his empirical assumption is false; if introspection instead is not a good way of forming beliefs about this matter, his empirical assumption is unjustified.

Be that as it may, the situation is even worse for the second component of Huemer's empirical assumption. For introspection is not even a prima facie promising way to form beliefs about causal relationships, even causal relationships between our own mental states. It has been recognized for a long time that we routinely, in particular, misidentify what reasons our beliefs are based on (see Harman [1986], particularly ch. 4, and references there).

One can emphasize the extent of the problem for Huemer's view by comparing PC with EC, Harman's conservative view described in 1.1. Assuming the first part of Huemer's assumption is correct, and every belief is accompanied by a seeming, EC is extensionally equivalent to PC; hence EC avoids Huemer's self-defeat argument, since every belief that is prima facie justified according to PC is also prima justified according to EC. However, EC has the advantage of avoiding self-defeat whether or not the second part of Huemer's controversial empirical assumption is true.

Interestingly, even granting extensional equivalence, we could still discriminate between the two forms of conservatism by thinking of possible beings which have beliefs, perceptions, memories, reasoning capacities, just like ours, but, at least *on some occasion*, no corresponding seemings. Clearly, they would be justified in their beliefs if and only if we are

(of course, that is partly because I believe I am one of such beings; but not only because of that), and PC cannot explain that.

Let us move to a second argument for the view that seemings, in the restricted sense discussed in section 1.2, provide justification. I said there that dispositions to believe that we are aware of in an indirect manner do not count as seemings. The latter point however opens the space to a new objection. The objector would first claim that dispositions to believe that a subject comes to know about indirectly, as in the example of a neuroscientist telling the subject about it, do not count as evidence, while "felt" dispositions, or seemings, do. To make this more precise: in the former case, your evidence includes the propositions that you have a disposition to believe *p*; this is not in itself evidence for *p*, although it could be if you had a reason to believe your dispositions in general, or this one in particular, to be a reliable guide to truth. But when the subject has a felt inclination to believe *p* (or, as the objector is likely to prefer saying, it seems to the subject that *p*), the subject's mental state in itself counts as evidence for *p*. If this step is granted, the objector will then have a good point; what does the evidential work is not the awareness of the disposition, but rather its peculiar phenomenological character. So after all my story just is a notational variant of something like Bealer's story, for there is a mental kind which plays a crucial epistemological role: feeling an inclination to believe.

Now I think the last step is dubious; for "feeling an inclination to believe" is certainly not a kind in the sense that it is fundamental, irreducible element of our mental constitution. If one can provide an explanation of how this composite kind can play an evidential role, than Bealer's claim is not completely vindicated. At this point, however, what I want to resist is the starting point of the argument, that seemings play an epistemic role which is not identical to that of an awareness of a disposition to believe. We noticed that there is little pre-theoretical pressure to take a seeming to be evidence, except insofar as we are aware of seemings in some area to be reliable indicators of the truth of their content. However, a case could be made that we still would like to make room for inclinations to believe to play some role in the formation of belief. A nice illustration of the problem is given by an example which Williamson devised; I will first illustrate the context in which the example is used and then I will consider whether the example might pose a problem for Williamson's own theory, and more generally for any theory on which seemings do not have an independent evidential role.

The example comes up in connection with criticism of the use, in accounting for the epistemic role of "intuitions", of a form of EC: one has a defeasible right to one's beliefs,

which may be defeated by positive reasons for doubt, but not by mere absence of independent justification.

Williamson's first point against using EC to illuminate philosophical methodology, is that it fails to provide an account of much we are interested in. I will not discuss how deep this criticism goes here. What I am interested in is his further objection, instead based on the incapacity of the account to make room for the evidential role of inclinations to believe. He provides the following case:

> Justin has been brought up to believe that knowledge is equivalent to justified true belief. He is confronted for the first time with a Gettier case. He might have immediately and confidently judged that the subject has justified true belief without knowledge, and abandoned his old belief that knowledge is equivalent to justified true belief. (…) Instead, Justin is more cautious, not wanting to assent too readily to anything tricky. Although he is consciously inclined to judge that the subject has justified true belief without knowledge, he does not immediately give in to that inclination or abandon his ingrained belief that knowledge is equivalent to justified true belief. (Williamson [2007] pp. 242-3)

In such a case, EC apparently does not give the prediction that Justin should abandon his belief that all knowledge is justified true belief. However, the defender of seemings as a source of evidence could observe, it is not clear how any other theory can fare any better, unless it just gives Justin's seeming an evidential role. Consider e.g. Williamson's own account, on which our evidence is just our knowledge. Clearly Justin has no knowledge that the subject in the Gettier case is a counterexample to the justified true belief analysis, since knowledge requires belief, but Justin does not believe that. Although he has no knowledge of the justified true belief account either, since we are supposing it is false, he might as well have defeasible evidence in its favour, as it provided the right verdict on a vast number of cases so far, and it is not clear (at least it is not clear to Justin) what could be substituted for it. Propositions that Justin knows thus support the traditional account inductively and abductively. The only thing Justin knows that tells against the traditional account is that he has an inclination to believe there is a counterexample; but that is not strong evidence[25]. Nor is Williamson here in a position to accuse Justin to give in to some generalized form of

[25] It has been suggested to me that perhaps Justin is in a position to use as evidence the fact that the Gettier case *might be* a counterexample to the justified true belief account. In general it would seem hasty to reject a theory because there is something which *might be* a counterexample. Most relevantly, on a standard semantics for the epistemic 'might', Justin can know that the case might be a counterexample only if he knows that he does not know that the justified true belief account holds. But he does not seem in a position to know that. So Williamson cannot allow the belief that the case might be a counterexample in Justin's evidence.

scepticism on judgment; there is no single actual judgment he is rejecting, and he is retaining trust in his everyday capacity for epistemic evaluations.

I think there is a reply open to Williamson (which is open to the epistemic conservative as well). After all, Justin is not now rationally obliged to change his mind. He might be, if he had formed the relevant belief, because then his evidential state would be different. Given the way things are, he is in fact rational to refrain from abandoning the traditional account. This does not imply that he is forever committed to an implausible theory. He might keep thinking about the matter; his inclination to believe will probably prove persistent, and it will extend to a number of relevantly similar cases. It's very important after all that we can produce Gettier cases with a certain ease. Suppose people generally had the intuition in the case originally described by Gettier but no one could think of a different (apparent) counterexample to the justified true belief theory of knowledge. My sociological prediction is that there would be no consensus that we found a genuine counterexample, and many theorists would try to explain away the case somehow. And it is at least not clear this would be an irrational stance to take. To use an analogy, when one observation provides a *prima facie* counterexample to a well established scientific theory, what scientists actually do is usually not giving up the theory, but rather looking at ways to contest the observation or explain it away. Naïve falsificationism is a bad account of how Science it's done, as even Popper admitted.

But as it happens, Justin could easily come up with new Gettier style examples, and the inclination to believe in those cases will prove solid as well. These new facts will somehow require an explanation, putting pressure on the old view. He might also find that his inclination to believe is in accordance with some plausible general principle, such that knowledge is incompatible with luck. This is not, on the whole, an implausible account of how many philosophers, with the best intellectual honesty, actually come to change their mind about theses they held dearly under the pressure of "intuitive" counterexamples.

Still, the foregoing is not complete as an explanation of what is going on in these cases. We also would like to have an account of why it would be right for one, at least *ceteris paribus*, to give the negative verdict on the presence of knowledge in a Gettier case. To see that, consider Justin's sister, Prudence, who has no philosophical education, and is not concerned with the justified true belief account of knowledge at all. Being presented with the Gettier case, she has an inclination to believe that the subject does not know. Yet, since she is just generally extremely cautious in judging things, she refrains from forming the judgment. Now we want to say that there is something wrong with Prudence. She does not, as Justin,

have other justified beliefs which foster her cautiousness. So, one might want to argue, she is somehow irrational in not forming the judgment. In particular the defender of the evidential role of seemings will want to say that she is failing to believe according to the evidence, which consists, in part, of the fact that it seems to her the subject does not know.

It might be suggested that the subject in cases of this sort might be irrational in failing to gather evidence which is open to her to easily gather, rather than in not respecting the evidence. I am not convinced though that the reply could handle this particular case. Prudence, we stipulate, has no particular interest in doing Epistemology. What is the practical reason she should form a judgement in this case?

I think a better reply is the following. We already noticed that knowing you have a disposition to believe something is evidence for it if you have justification to believe that your dispositions are a reliable guide to truth in that field. In normal circumstances, we presuppose our conceptual faculties are working fine. If someone describes an object, saying it is used to drink, it has a cylindrical shape, and so on, or if one shows it to me, I will judge that it is a glass. What warrants my judgement here is that I have a competence in using the concept GLASS, which derives both from experience with actual glasses and from my knowledge of the language, and I reach the judgment by employing such a competence. Similarly with the Gettier case, on the picture Williamson gives, which I find congenial, and which I will try to develop in chapter 4, our disposition to believe stems from our competence with an ordinary application of the an ordinary concept[26]. The point here is not being reliable in the application of the concept makes you justified in the transition between inclinations to believe and beliefs; since the relevant belief-forming process, in the cases we are imagining, has not been completed, its reliability cannot have normative bearing. The point is one about rational tension among beliefs. Suppose that if the cylindrical object were described to Prudence, she would refrain from judging it to be a glass, while being aware of her inclination to believe that. This would conflict with her belief that she usually applies the concept GLASS with competence (or, as the non-philosopher is more likely to say, her belief that she knows what a glass is, or her belief that she knows what 'glass' means). Other things being equal, it would improve her epistemic situation to remove this conflict. Suppose instead she has good reasons to doubt that she is applying the relevant concept with competence; we might imagine, e.g., she is participating in a Psychology experiment such that she had a fifty per cent chance of

[26] While it is hard to deny the concepts in question are very often ordinary ones (like KNOWLEDGE), it might be thought the cases that are considered in philosophy are out of the ordinary or far-fetched in some problematic ways. I believe this is not usually the case; I will come back to this matter in chapter 5.

being given a pill which would completely impair her conceptual capacities, while leaving her an illusory sense of lucidity.

It is a further question, which perhaps a sceptic could press, whether we generally have any better justification for crediting ourselves with a reliable capacity to apply our concepts. Barring this doubt, the view that seemings are felt dispositions to believe has no problem in accounting even for this last case.

Conclusion

I have considered two broad forms of argument to posit intuition as a *sui generis* mental state. The first relied on epistemic considerations to argue that any position which rejects intuitions as a mental state which plays an important epistemic role is incoherent. But the argument seemed to vastly overlook the resources available to the defender of such a position.

The second sort of argument identified intuitions with seemings, while claiming the latter constitute a mental state distinct from belief and not analyzable in terms of belief. I offered an account of seemings in terms of felt inclinations to believe. I argued that while seemings in this sense are distinct from belief, and one could use the term "intuition" to denote them, there is no reason to claim that seemings, so understood, play a central epistemic role.

Chapter 2
Conceptual Analysis under Scrutiny

We do not learn first what to talk about and then what to say about it
Willard Van Orman Quine

Berkeley thought that chairs are mental*, for Heaven's sake!*
Jerry Fodor

Introduction

Chapter 1 talked about one way in which philosophy, and philosophical thought experiments in particular, are thought to be epistemologically special, that is, by relying centrally on intuitions or seemings, and rejected that idea. Sometimes the picture of philosophy on which it relies crucially on intuitions, understood as a special sort of mental state, is coupled with the claim that central to philosophy is an activity called 'conceptual analysis'. The latter claim is still, I believe, fairly common both among philosophers and among non-philosophers (those who have an opinion on the matter). The meaning of the expression 'conceptual analysis' is vague enough for me to agree with that claim, on some (fairly uninteresting) interpretation. I believe, for example, that it is fairly central to philosophy to devise thought experiments in order to provide counterexamples to certain theories or definitions. But I will argue that on some more interesting interpretation the claim that philosophers engage in 'conceptual analysis' is not correct. Section 2.1 tries to spell out an interesting meaning of 'conceptual analysis'. On such reading, conceptual analysis is characterized by the fact that judgments about thought experiments have some special epistemic property, which derives in turn from a special link between those judgements and the concepts involved. I then discuss some ways of spelling out the alleged link between the

judgments and the concepts involved, and some problems for the resulting views. In section 2.2 I introduce a specific view, epistemic two-dimensionalism, on which there is a level of content, primary content, which serves two roles: being internalistic, in the sense of supervening on features of the individual that grasps the content, and grounding the privileged epistemic status of a class of (ideal) judgements about possible cases. The latter result is achieved by defining primary content as a function of such judgments. One could then understand conceptual analysis as the business of finding truths that are grounded in primary content. Section 2.3 introduces some thought experiments devised by Burge to argue that some terms do not have primary content in the sense just explained; I will discuss these arguments in the light of epistemic two-dimensionalism, and I will consider the way Chalmers, a prominent defender of two-dimensionalism, has proposed to deal with Burge's cases. I will argue that Chalmers' reply is unsatisfactory and that the arguments provide conclusive reasons against the idea of primary content, without begging any question against two-dimensionalism. In section 2.4 I turn to the idea of defining some kind content in terms of ideal judgments about possible cases, and I conclude that, in the light of the rejection of internalistic content, the idealization involved makes the definition either empty or unmotivated.

2.1. Analyticity and Conceptual Analysis

One way in which philosophical thought experiments are sometimes thought to be special, is by being at the heart of a practice called "conceptual analysis"[27]. One, admittedly rough, characterization of such a practice is the following; we consider imaginary cases, and we judge whether a certain concept would apply in such a case. The purpose of such an activity is to shed light on the concept itself. This is possible because these judgements are constitutively tied to the concepts in some way; different views will differ on the exact nature of such a tie. However, on any view of conceptual analysis worth the name, the link between the judgment and the concept cannot be exhausted by a tie between the identity of the concepts and the truth-value of the content of the judgment (that would be trivial; identity of intension is at least necessary for identity of concepts). The judgment itself (the mental act),

[27] The notion of conceptual analysis that I will talk about is not to be confused with the notion employed by some authors influenced by the later Wittgenstein; see e.g. Strawson [1992], in particular chs. 1-3. In the sense of these authors "conceptual analysis", as far as I can understand, is a description of the relations existing among our concepts as we actually use them; therefore, it does not try to discover analytic or necessary truths, but rather contingent empirical truths about our "conceptual scheme".

and not just its content, needs to be tied to the possession of the concepts, either psychologically or epistemically[28].

As I said, the foregoing characterization is sort of rough, and I will consider various possible ways of developing it. In particular, the relation between the relevant judgments and the possession conditions of the concepts can be spelled out in various ways. However, the picture just sketched is a very influential one; its crucial elements are clearly recognizable e.g. in Bealer [1998], Bonjour [1998], Jackson [1998], Goldman and Pust [1998], Pust [2000], Peacocke [2000], Chalmers and Jackson [2001], Henderson and Horgan [2001], Chalmers [2002], Goldman [2007], Ludwig [2010], Nimtz [2010].

One natural way of spelling out the idea of a link between certain judgements and the possession of a certain concept is saying that (justifiably) accepting these judgements is *necessary* to possess the concept involved. In the terminology of Boghossian [1996] the fact that (justifiably) accepting a judgement is necessary to possess the concepts involved means that it is "epistemically analytic". The notion of *epistemical* analyticity contrasts with the notion of *metaphysical* analyticity. A judgment is metaphysically analytic when it is true in virtue of meaning only[29]. While Boghossian thinks that epistemical analyticity is a useful notion and it plays a crucial role in the epistemology of the a priori (see Boghossian [1996], [2003a], [2003b]), he is not committed to the picture of conceptual analysis I just sketched. Ludwig [2010] seems to endorse conceptual analysis together with epistemic analyticity, where he identifies philosophical "intuitions" with judgements that are "*solely* an expression of one's competence in the contained concepts and their mode of combination." (p. 433, original emphasis). Ludwig makes clear that the relation between the competence and the judgement is meant to be etiological. He does not take, strictly speaking, the possession of the concepts to be sufficient for the belief, but he seems to think that considering the relevant proposition is the only further condition that needs to be satisfied. Clearly, the notion of

[28] For a defense of a theory of concepts that sits well with this approach, see in particular Peacocke [1992]. Of course one could doubt that investigating possession conditions is in general a good approach to the investigation of concepts. I heard Jerry Fodor saying (although I could not find exactly the same idea expressed in print), that it makes just as much sense to investigate elephants starting from the question: 'what does it take to possess an elephant?'. Nothing here will depend on the issue anyway.

[29] I will not talk in this work about metaphysical analyticity, since the focus of this work is more on the epistemology of philosophy than on its subject matter. Of course, it is controversial whether there is a useful notion of truth in virtue of meaning; Williamson [2007] argues to the contrary. See Russell [2008] and Stalnaker [2011] for a defense of the notion, but also Boghossian [2011a] and Williamson [2011b] for criticisms of these defenses.

epistemical analyticity has an independent interest; I will discuss in the next chapter some epistemological consequences of giving up the notion. Moreover, and most importantly at this point, it provides a particularly clear way of filling in the details in the picture of conceptual analysis described above. As we will see, various other ways of providing these details can be naturally described as modifications of this idea.

The main problem with the idea of epistemically analytic judgements is that it is very hard to find enough of them. Williamson [2007] argues that there are none to be found. Note that in order for epistemic analyticity to play the crucial role in the broad characterization of conceptual analysis offered above, more than an existential claim would be needed. Suppose, for the sake of the argument, that every instance of the law of non-contradiction (anything of the form: not-(P and not-P)) is epistemically analytic, and nothing else is. Then conceptual analysis as I described it would not be either an adequate representation of the way philosophy is practiced, or a useful intellectual practice on its own. We can think of imaginary cases and conclude that the subjects in those cases do not at the same time know and fail to know a certain proposition; but clearly the epistemological interest of these "thought experiments" is very limited. So while reviewing Williamson's argument against epistemic analyticity, and various possible replies, we need to keep in mind that even if one is not convinced that Williamson established that there are no epistemically analytic truths, one might be convinced that they are not numerous enough, and of the right kind, to ground the practice of conceptual analysis.

I will not rehearse Williamson's arguments in their entirety, but I will try to give the gist of them. In Boghossian's initial characterization a statement[30] is epistemically analytic just in case "grasp of its meaning alone suffices for justified belief in its truth"[31]. This characterization is quite obviously inadequate, since it does not take into account the fact that we humans are prone to error. Instances of logical truths, at least simple ones, are supposed to be paradigmatic cases of epistemically analytic truths; but people who are capable of understanding them can make mistakes, even trivial ones, and thereby can fail to have justified beliefs in many particular instances of trivial logical truth, by failing to have belief at

[30] Throughout this chapter, I will switch freely between talking about words and sentences, and talking about concepts and thoughts; unless otherwise specified I will take what I say about the ones to apply to the others and viceversa. Of course this is not unproblematic. However, the philosophers who are targeted here tend to do the same, and, most importantly, my view is that the arguments I will consider have the same force with respect to linguistic and mental content.

[31] Boghossian [1996] p. 363

all. Boghossian seems to be unaware of the problem in his [1996]. In later writings, the formulation he gives is more guarded, but also less clear; in [2003b] he writes that a sentence is epistemically analytic "if grasp of its meaning *can* suffice for justified belief in the truth of the proposition it expresses" (Boghossian [2003b], p. 15, emphasis added). I suppose something *can* suffice for something else when it suffices *provided that* some other conditions are satisfied, but Boghossian never explains what these conditions are[32].

There are at least a couple of simple ways in which defenders of something close to epistemical analyticity might try to overcome the problem. We might, on one hand, qualify the sort of grasp that is needed, restricting the definition to *full* grasp (see Peacocke [1992] and Bealer [1998]). On the other hand, we might qualify the sort of belief or assent that the statement must compel, for example by saying that the subject who grasps the statement must be *disposed* to believe it; and we might also combine the qualifications, saying that a statement is epistemically analytic if someone who fully grasps it has a disposition to believe it.

A counterexample to the characterization of analyticity in terms of a link between understanding and disposition to assent can be extracted by some cases devised by Burge in a different context. I will be very brief here both in the description of the case and in its discussion, because I will come back to such cases at great length in the subsequent sections of the chapter. Consider Bert, who has suffered from arthritis for many years, and is (seemingly) aware of this, but ignores that by definition arthritis can only occur in the joints. The sentence "I do not have arthritis in my thigh" is a very plausible candidate for analytic truth. But Bert, at some point, comes to think he does have arthritis in his thigh. At that point, he seems to understand the supposedly analytic sentence, but, absent the gathering of new information, he seems to have no disposition at all to accept it.

Of course, the foregoing example does not tell against the specification of analyticity in terms of a link between *full* understanding and (disposition to) assent. That qualification is proved insufficient by a further kind of counterexamples, those of *experts* with a stable disposition to reject a paradigmatic (putative) epistemically analytic truth. Of course the

[32] By 2011, Boghossian writes, in discussing Williamson's objections: "we are already aware, from reflection on the paradox of analysis, that analytic truths do not have to be transparent. If correct analyses can sometimes be informative, that must be because they don't always seem correct. So we can't conclude from the fact that a person competent with word w denies S(w) that S(w) is not analytic for that person." (Boghossian [2011] p. 492). But this seems to straightforwardly contradict his 1996 definition of analyticity. I surmise Boghossian was not occurrently aware of the problem posed by the paradox of analysis at that point.

possibility of cases like that would suffice to yield an objection to the picture, but we happen to have actual cases. Williamson invites us to consider Vann McGee, who presented (McGee [1985]) some putative counterexamples to the validity of *modus ponens*, which Boghossian proposed as a paradigmatic example of an epistemically analytic rule of inference (to which, I am assuming, some logical truths will correspond, derived from mere application of the rule). Here is the most well-known of McGee's cases: suppose we are in 1980, and there are three candidates for the American presidential elections. Two of them are republicans, Reagan and Anderson, and one is a democrat, Carter. The polls show Reagan at 60%, Carter at 39% and Anderson at 1%. In the past history, no candidate with an advantage in the polls similar to the one Reagan has ever lost. So it's rational for someone in this situation to believe: P) Reagan will win the elections. P entails

P1) A republican will win the election

It is also rational, given the situation, to believe

P2) If a republican will win the election, then, if Reagan does not win, then Anderson will

P1 and P2 entail, by *modus ponens*,

P3) If Reagan does not win, then Anderson will

But P3 is, or seems to be, clearly false. If Reagan does not win, Carter will, and so Anderson will not.

McGee thinks that someone in that situation is correct in accepting P1 and P2 and rejecting P3. So he has a stable disposition, after careful reflection, to reject, given contexts of the appropriate kind, an instance of a simple logical truth (e.g. the conditional having the conjunction of P1 and P2 as an antecedent and P3 as a consequent). As a matter of fact, not only he does not have a disposition to accept it, but it seems most of us do not have it, for the counterexample is cleverly constructed in such a way to have some initial attraction. Moreover, McGee's case is very different from paradigmatic cases of 'partial understanding', like the one discussed in Burge [1979] which we have mentioned above. McGee is not willing to defer to anyone on the correct meaning of 'if', and he is an expert on the use of conditionals, by any socially recognized standard. It is important to see that for McGee's example to be a problematic case for the definition of epistemic analyticity it is not required that it is convincing (in fact, we are presupposing it is ultimately incorrect) or even worth of the serious consideration it has in fact received. It is only required that McGee's argument is not so absurd that we are inclined to deny he could even engage in the discussion, since he does not use the word "if" with the same meaning we do.

Williamson also considers a number of imaginary cases, which seem to be equally convincing[33]. Boghossian [2011b] replies to Williamson trying to defend the existential claim that there are epistemically analytic judgements, but he concentrates not on his original example of *modus ponens*, but rather on the conjunction-elimination rule. However, as we observed before, the existential claim is really not interesting in the present context; surely we do not reach interesting philosophical conclusions merely by conjunction-elimination (nor does Boghossian suggests otherwise). The general moral of the discussion, for our purposes, is that for most concepts at least, for any given specification of possession conditions in terms of judgements or inferences that the subject is disposed to accept, we can either find of coherently conceive of someone who a) does not satisfy the possession condition and b) is plausibly regarded as fully possessing the concept.

I am not going to argue in general against any possible way to recover a traditional idea of conceptual analysis. Arguing at that very abstract level, it seems to me that Williamson has made a good case that no such idea will survive. But we know that proving negative existentials is very hard, and it is particularly hard when what should be proved not to exist is very loosely defined. What I supply in the next sections, instead, is some detailed criticism of one specific view which is prominent among contemporary defenders of conceptual analysis, and was absent from Williamson's focus. If such criticisms are successful of course the general arguments considered here will be reinforced.

The theory I am going to concentrate on for the rest of chapter, epistemic two-dimensionalism, has a sufficiently broad application to ground an epistemology of the a priori, and of (philosophical) thought experiments. On that theory, as we will better see, one also recovers a notion very similar in spirit to epistemic analyticity; a judgement is said to

[33] The central imaginary cases devised and discussed by Williamson concern the sentence "every vixen is a vixen". This is a modification of 'every vixen is a female fox', an example mentioned by Putnam in his [1962]. Putnam deemed the sentence to be a clear, although uninteresting, case of analyticity; his view in that paper is that there is a sensible analytic-synthetic distinction, but it has extremely little epistemological or theoretical usefulness. Interestingly, Quine [1960] refers to Putnam's paper, then in print, approvingly (p. 57), and he seems to have held a similar view of analyticity from that point of his career on (see Hylton [2007], ch. 3). Considerations of the same kind that support semantic externalism make it utterly implausible that 'every vixen is a female fox' is epistemically analytic. One could acquire e.g. the concept 'vixen' by ostension and the concept 'fox' by testimony without realizing that their denotations overlap. Williamson sets his bar higher, by using 'every vixen is a vixen' as an example; in arguing that even the latter sentence fails to be analytic, he is a far more radical critic of the analytic-synthetic distinction (or, at least, of one such distinction) than Putnam, or, indeed, Quine.

express a conceptual truth just in case an ideally rational version of the subject who entertains the judgment would accept it under any circumstance.

A further reason of interest of the following sections should be the following. I will explore the links between the idea of epistemical analyticity (or conceptual truth), and a certain view of the nature of content, the internalist view. On the internalist view, at least some level of intentional content of any mental state or linguistic act supervenes on the internal state of the individual which is in that state or performs that act. I will argue that the idea of a kind of content which supervenes on the individual is behind at least the specific defense of conceptual analysis discussed in this chapter, in the sense that if one gives up internalism, one has to give up that picture as well.

2.2. Twin Earth and Two-Dimensionalism

Let us consider the following two theses:

1. There is a primary level of content which supervenes on the total internal state of the subject entertaining the content
2. There is a primary level of content which supervenes on judgements about possible cases under idealized rational reflection (i.e. putting aside confusion, lack of time, and so on).

Just a few initial clarifications about 1 and 2; 'Content' is here taken to stand for any truth-apt entity or state a human being can grasp, or be in (see also note 3). Where it is said that there is a certain *primary* level content, it is to be understood that there might be other levels of content, but also that whenever content of a different kind is grasped, content of the primary level is also grasped. The view I am going to discuss, two-dimensionalism as endorsed by Franck Jackson and David Chalmers[34], or epistemic two-dimensionalism (henceforth, 'two-dimensionalism' will refer exclusively to this kind of view[35]) includes both 1 and 2, and identifies the content talked about in the two theses in one kind. The real target of my criticisms here, as I will explain momentarily, is 2. But I will argue that, if 1 is denied, 2 loses any initial motivation it might have. Therefore, I will spend quite a lot of time criticising 1.

[34] See Jackson [1998], Chalmers and Jackson [2001], Chalmers [2002], [2006a], [2006b].

[35] For discussion of alternative interpretations of the two-dimensionalist formal framework, and the defense of a perspective in accordance with the present work, see Stalnaker [2001] and [2004].

The view can be seen, as I anticipated, as a radical way of remedying the shortcomings of the definition of epistemic analyticity (although its proponents do not use the same terminology, I do not believe they could substantially disagree) we discussed in previous section, by saying that a statement is epistemically analytic just in case grasp of its meaning is sufficient for *an idealized version of* the subject to judge it to be true. This move follows in fact from the fact that the primary content of a statement is defined, as we will see in more detail below, in terms of the situations in which an idealized subject would judge it to be true. If the theory works, it promises to yield a recipe for individuating possession conditions in terms of epistemically analytic judgements for any concept whatsoever. Needless to say, if such a theory were successful, it would constitute an impressive achievement. It would establish a rather radical form of internalism. Often externalism is stated as the thesis that some content fails to supervene on the internal state of the individual; call that content *wide*. If internalism is the negation of externalism, it is then the extremely ambitious universal claim that no content whatsoever is wide. Two-dimensionalists are not strictly speaking internalists in this sense, as we will see, because they admit some wide content; but they come close, for they make the universal claim that whenever wide content is grasped, this is only made possible by the grasping of a closely related non-wide, primary content. Therefore, beyond its implications for conceptual analysis and the epistemology of thought experiments, two-dimensionalism, if established, would save a powerful picture, one that was extremely influential at least in the 20th century, of the fundamental nature of content. But I am going to argue that this result has not been achieved.

Let us start looking at how two-dimensional semantics handles a well-known case which some (see e.g. Burge [1982a], [1982b]) believe to constitute a counterexample to 1, Putnam's Twin Earth thought experiment. Looking at this thought experiment will serve the purpose of clarifying the content of 1 and, more generally, the content of the two-dimensionalist view. We can distinguish two versions of the Twin Earth thought experiment. In the original version, we are asked to imagine there is, in our universe, a distant planet, call it Twin Earth (TE), almost perfectly identical to Earth, containing individuals relevantly similar to us, except that in TE there is some stuff superficially identical to water, and playing the same role in TE's inhabitants' lives as water plays in ours, which is not H2O but rather XYZ. We can then consider two individuals, called Oscar and TwinOscar, who are relevantly similar[36] in their physical and phenomenological states; they both utter sentences containing

[36] Not exactly identical, because of course human beings are made in large part of H2O; if you think this confuses the thought experiment you can run it using a different natural kind term, e.g. 'aluminium', or 'dog'.

the (syntactically individuated) word ‘water’. Oscar and TwinOscar however, according to the objection against 1, mean something different by ‘water’; Oscar uses a word that refers to H2O, and TwinOscar a word that refers to XYZ. But this difference is not based on any difference in their internal physical properties. If there are mental properties which do not supervene on physical states, and can be described non-intentionally, we will count those as internal as well. It is generally agreed that Oscar and TwinOscar do not differ in any relevant internal property.

The second version of the thought experiment differs from the first only in that TwinEarth is not supposed to be a distant planet, but rather a metaphysically possible world. Our word ‘water’, we then notice, does not apply to the stuff in TwinEarth; in this version too, Oscar and TwinOscar, although internally similar in all relevant aspects, have two different (semantically individuated) words. Therefore, the content of the words does not supervene on their internal states.

The idea at the basis of two-dimensional semantics is conceding that the extension of the words used by Oscar and TwinOscar differ, while recovering one level of content that is shared by the two subjects. The content will be associated with an expression-token, a particular use of an expression. Let’s call the content Oscar and TwinOscar uses of ‘water’ share ‘primary intension’. The primary intension will be an identical function from centered possible worlds[37] (that is possible worlds with a privileged speaker and time) to extensions. The expression token is also associated to a secondary intension. Secondary intensions will be the more familiar functions from possible worlds to extensions. Since Oscar and TwinOscar inhabit different centered possible worlds, although the primary intension of their use of ‘water’ is the same, it determines different secondary intension; Oscar’s usage picks out H2O at all possible worlds, and TwinOscar’s XYZ.

It is useful, in order to understand the primary intension of ‘water’, to think of it as roughly expressed by a description (although we shouldn’t, at least in Chalmers’ version of the theory, identify the two) of the superficial features of water, which we might abbreviate as ‘the waterish stuff around here’. The ‘around here’ part of the description has to be added on account of the original version of the TE thought experiment. If we were to think just of the other version, a description like ‘the actual waterish stuff’ would be equally effective; but if we consider the distant planet version, the difference between the reference of Oscar’s and TwinOscar’s words would be left unexplained.

[37] There are other possible ways to understand the “worlds” primary intensions work on; I don’t think the differences matter for the following. See Chalmers [2006a] for an overview of the different possibilities here.

Moreover, crucially for my present purposes, the primary intension is supposed to ground a priori judgements. An occurrence of a sentence is a priori if and only if its primary intension is true at all centered possible worlds. Claim 2 above fits into this picture because the extension of a primary intension at a centered possible world is identified with what the (actual world) speaker would judge that extension to be under idealized rational reflection, given a complete description of the relevant centered possible world stated without using the expression in question. Conceptual analysis, on this picture, requires just consideration of various possible centered possible worlds, and, provided we approximate well enough the ideal rational reasoner, it delivers truths which are purely grounded in their own content (Jackson [1998] and Chalmers and Jackson [2001] are most explicit about the connection).

Two-dimensionalism thus offers a *prima facie* consistent strategy for holding 1 when confronted with Putnam's thought experiment. Such a strategy has of course been criticized in a number of ways (see Chalmers [2006b] for a survey of objections and replies). However, the different kind of thought experiment that I will consider poses, I will argue, a more radical threat to 1 than the TE case. The kind of thought experiments I am thinking about has been devised by Burge (see in particular Burge [1979] and [1986]). As far as I know, discussion of the problem these thought experiments pose is mostly absent in the vast literature on two-dimensionalism[38,39]. In the next section, which will be the more substantial, I will argue that the response offered so far by Chalmers does not succeed. In the last section I will argue, firstly, that the negation of 1, with some uncontroversial assumptions, entails the negation of

[38] Apart for the exceptions that are going to be noticed (in particular Chalmers [2002] and Schroeter [2005]), there is a problem sometimes discussed in the literature about the supposed aprioricity of the features determined by primary intensions, which resembles the one I am going to discuss. Chalmers [2006b] discusses the problem in the version proposed by Block and Stalnaker [1999]. See also Laurence and Margolis [2003] for a version of this sort of criticism directed at Jackson [1998], and Owens [2003] for a somewhat analogous use of Burge's thought experiments with respect to Kaplanian characters rather than primary intensions.

[39] Some have argued that the sort of externalism that derives from accepting the thought experiments I am going to use against 1 and 2 is incompatible with the idea that rational subjects have a priori knowlegde of, or privileged access to, the contents of their own thoughts. This view is argued for in various ways, e.g. in McKinsey [1991] and [2007], Brown [1995] and [2001], and Boghossian [1989] and [1998]. The opposite view, defending the compatibility of externalism and privileged access, is defended e.g. in Brueckner [1992], Warfield [1992], McLaughlin and Tye [1998], Goldberg [2003a] and [2003b], and Brown [2004]. My arguments here are fully consistent with both positions, as far as I can see. If incompatibilism is true, and one still maintains semantic anti-individualism, then perhaps one is in a position to provide an independent argument against two-dimensionalism. The only assumption I need here is that the compatibility problem does not provide grounds for rejecting semantic anti-individualism. This is not uncontroversial, but it is by far the prevailing view.

2, on any plausible reading of 2. Secondly, I will argue that 1 cannot be defended by appeal to 2.

2.3. Thinking about Arthritis and 'Arthritis'

2.3.1. Enter Arthritis

The only general theoretical assumption we need to set up the thought experiments, in order to use them as a counter-examples to the two-dimensionalist defence of 1, is the following: it is possible to believe something which is a priori[40] false (by which I mean – standardly – that it would be possible for someone to know a priori the negation of the belief). This is not at all something the two-dimensionalist wants to deny. In fact, I take it to be part of the view that there are such beliefs, and therefore the argument works, *ad hominem*, even if there can be no a priori false beliefs, e.g. because a priori knowledge is not possible. To make vivid the possibility required by the assumption, consider the case of mathematical beliefs; in a two-dimensionalist framework, the primary intension of '1478+355=1838' makes it the case that the statement is a priori false, but it seems we can believe it nonetheless. We can just, e.g., go wrong on the calculation, or rely on misleading testimony. It would be quite incredible to claim that someone going wrong in the calculation in the described way has a different concept of +, or =, or of the numbers involved. Of course the two-dimensionalist would advance a similar claim if the subject were to persist in her mistake under idealized rational reflection; but this is not the case we are imagining here. Obviously, more complicated examples could be provided, such as the possibility of believing the negation of Fermat's theorem.

Let me also clarify two assumptions the thought experiments does *not* rest on. First, the thought experiment does not rely on the claim that we cannot, when faced with extravagant error, reinterpret the words of a speaker non-homophonically; if I were to encounter a speaker saying 'arthritis is a form of Japanese theatre', I should not attribute that speaker any belief about arthritis. So, if I attribute her any belief at all, I should do it non-homophonically. What is instead assumed is that some of the time the correct thing to do is attributing false beliefs by interpreting homophonically false utterances. Moreover, the

[40] As a matter of fact, I am going to argue in next chapter that the notion of a priori is too unclear to do serious epistemological work. At this stage, however, I am conceding the two-dimensionalist that the notion is meaningful. So, more precisely, my assumption is that it is possible to believe something which is very plausibly a priori false if one endorses the two-dimensionalist framework.

thought experiments will *not* rest on the claim that content has to be individuated through appeal to a public language; we might talk merely about the subject's idiolect. Since the possibility of some variation between different individuals' idiolects is built into the two-dimensional framework, it would beg the question to assume that we just speak a public language. But again, I do not assume anything of the like. The thought experiment is concerned with the content of the subject's thoughts and utterances; sometimes the subject's interaction with other subjects will play an evidential role, but based mostly on particular judgements and through no general hidden assumptions.

Now consider again Bert, who believes (or so it seems) he has arthritis at his thigh. We can plausibly start with the assumption that the content of his belief, as we normally understand it, is something which is a priori false, because arthritis is by definition an inflammation of the joints, and a thigh is not a joint[41]; and in the two-dimensionalist framework this means it is false because of the primary intension of the terms involved. Now, one move that I would like to put aside, for the time being, is claiming Bert's belief is not false after all (Chalmers, as we will see, does not take this route). Since it's just one case we need, we can set up the details of the thought experiment in such a way to make that move utterly implausible. Bert, if corrected, would not regard his previous belief as true, or meaningless, he would regard it as plainly false. If the belief was true, we would then be forced to attribute semantic blindness to Bert. Let us also imagine that Bert has had arthritis in his knee for many years, and we would naturally attribute him many true beliefs, without any need for reinterpretation, about arthritis: that he had it for many years, that it is an annoying disease, that it is the same disease his friend Adam and his father have, and so on. Applying the principle of charity to this one belief would then be completely unjustified (remember we already assumed you can believe a priori falsities). Before considering more possible reinterpretations, let us move to the second step of the thought experiment; it seems we can conceive easily someone, call him TwinBert, who is identical to Bert in internal properties, and went through the exact same life (non-intentionally described), but lives in a different

[41] Brian Weatherson pointed out to me that it is not clear that it is a priori (even in the two-dimensionalist framework) that a thigh is not a joint. If it were not, we would need to reformulate the discussion in terms of Bert's conjunctive believe that he has arthritis in his thigh and his thigh is not a joint. Such a belief can be a priori false even if the second conjunct is empirically known. So there would be no difference in the ensuing dialectics. For simplicity's sake, I will ignore this complication and ask the reader to go along with the assumption that it is a priori that a thigh is not a joint. Burge does not actually say that his thought experiments involve beliefs that are a priori false; he might think otherwise. Burge [1986] describes the more general problem as arising from beliefs in necessarily false propositions.

linguistic community. In TwinBert's linguistic community 'arthritis' can apply to whatever illness he and Bert both have at the thigh. So it seems TwinBert has a true belief, and that is not because of any difference in the relevant physical facts; so it must be because the content of his belief is different. But Bert and TwinBert are identical in their internal states, so the content of their beliefs does not supervene on their internal states. Moreover, and this is how the case differs from the TE case, if Bert's belief is false because of its primary intension, than the primary intension also fails to supervene on the subject's internal states.

2.3.2. The Meta-linguistic Move and its Problems

What the two-dimensionalist needs to argue at this point, if she is to parallel her move in the TE case, is that the difference in content between Bert's and TwinBert's thoughts is not a difference in primary intension. Chalmers [2002] argues exactly that, and the same strategy is briefly defended in Chalmers [forthcoming]. Let me stress that I think this is the best route to take for the two-dimensionalist, since, among other things, it would allow us to maintain that Bert's belief is false. Here is how Chalmers [2002] tries to achieve this result:

> Here, the crucial factor is that Bert uses the term 'arthritis' with *semantic deference*, intending (at least tacitly) to use the word for the same phenomenon for which others in the community use it. We might say that this term expresses a *deferential concept* for Bert: one whose extension depends on the way the corresponding term is used in a subject's linguistic community. It is clear that for deferential concepts, extension can depend on a subject's environment, as can subjunctive intension. The subjunctive intension of Bert's concept *arthritis* picks out arthritis in all worlds, while the subjunctive intension of Twin Bert's concept picks out twarthritis in all worlds. (...) In general, the epistemic intension of their *arthritis* concepts in those scenarios will pick out the extension of the term 'arthritis' as used in the linguistic community around the center of those scenarios. (Chalmers, [2002], p. 617, original italics)

This move is consistent, and ingenious, but the problem is whether it is plausible. The first thing to notice is that the concept Bert has will have to be in some sense meta-linguistic; the primary intension of 'arthritis' as used by Bert will have to be rendered by something like 'the phenomenon denoted by 'arthritis' in my linguistic community'[42].

Now, though, notice also that the motivation for attributing a similar non-standard concept to Bert is that he intends to "to use the word for the same phenomenon for which others in the community use it". However, this motivation does not only apply to Bert, it

[42] The reference to "my" community is needed to take care of the possible version of Burge's thought experiment involving a distant planet in the actual world, or just an unconnected linguistic community somewhere in this planet.

applies to most speakers; most people who correctly believe that arthritis can only appear in joints, would be disposed to change their mind if a sufficiently authoritative speaker, or maybe more than one, were to correct them. Their concept seems to be no more and no less deferential than Bert's one. I don't see how Chalmers can object to this extension; remember TwinBert, after all, is a competent speaker of his community, but he also has a deferential concept. Notice also that the same thought applies to an extremely wide class of concepts; Chalmers writes:

> One can apply the same reasoning to Putnam's case of 'elm' and 'beech', in which a subject can use the terms with different referents despite users having no substantive knowledge to differentiate the two. In this case, the terms are being used deferentially: the epistemic intension of the subject's concept *elm* picks out roughly whatever is called 'elm' around the center of a scenario, and the epistemic intension of her concept *beech* picks out roughly whatever is called 'beech' around the center of a scenario. (Chalmers [2002], p.618)

Burge [1979] famously provided a similar thought experiment for 'sofa'[43]. He adds the following plausible remark:

> 'Brisket', 'contract', 'recession', 'sonata', 'deer', 'elm' (to borrow a well known example), 'pre-amplifier', 'carburetor', 'gothic', 'fermentation', probably provide analogous cases. Continuing the list is largely a matter of patience.' (Burge, [1979], p. 86)

So both Chalmers and Burge agree that the relevant considerations apply over a vast number of terms belonging to many different categories. If we couple this point with the observation that the grounds for attributing what Chalmers calls a deferential concept are the same for speakers who do not make linguistic errors as for those who do, we start seeing that Chalmers' reply involves holding that the majority of concepts for the majority of people are of a meta-linguistic kind. Even these limitations are not clearly motivated; experts too, sometimes, defer to other experts. And if the rationale for attributing the deferential concept to Bert is that he intends, to quote Chalmers again, "to use the word for the same phenomenon for which others in the community use it", since this seems a pre-requisite of being part of a community of speakers, perhaps the reply involves thinking that *all* concepts of *any* rational subject who is part of a community of speakers are of a meta-linguistic kind.

Chalmers (p.c.) replied here that one could both use the term deferentially (and thus have the deferential concept) and use at other times the term non-deferentially (and thus have

[43] This case involves someone convinced that a large armchair can qualify as a sofa. The case of course should not be confused with the one described in the Burge [1986] that I am going to discuss later.

the non-deferential concept). I would concede that it is a logical possibility that we have both kinds of concepts. However, if the only motivation for positing the non-deferential concepts were avoiding the result that most of our concepts are deferential, it would be blatantly *ad hoc*. In other words, we need to have a reason to think that the non-deferential concepts are used at some point by the subject. But this seems to conflict with our considerations so far. The use of 'arthritis' which was singled out by Burge is in no way atypical or special, with respect to the subjects' intention to use it deferentially, and in no way particularly apt to be the object of the thought experiment. There is no reason to think the same sort of reasoning would not apply to all the other terms and concepts we mentioned in this connection.

That the result is so widespread might already be a reason for some concern. Whatever the intrinsic plausibility of the result, the difficulty for restricting the meta-linguistic move worsens the next problems for Chalmers' reply which I am going to present. I will discuss three objections (which I will indicate as A, B and C), and dismiss some possible replies. I regard each objection as posing a very serious problem, and the three together as leaving no hope that the move could be rescued with some adjustment.

A) A first unwelcome consequence of the meta-linguistic move is that it gives extremely counterintuitive judgments about the identities of meanings and concepts. The worry comes in various forms. First, and most obviously, the meta-linguistic account of primary intensions makes it impossible for words from languages other than English to mean the same as 'arthritis', and the other way round. For example, the primary intension of (utterances of) the Italian word 'artrite' will be roughly 'whatever is called 'artrite' in my linguistic community', which is different of course from 'whatever is called 'arthritis' in my linguistic community'. The point here does not depend in any way on taking seriously the description offered as a gloss to the primary intension; the English term and the Italian one determine different functions from possible worlds to secondary intensions. Here is why: surely it could be that in Italian 'artrite' applies outside of the joints, and 'arthritis' in English doesn't. It is crucial to Chalmer's solution that the syntactic form which is used differently with respect to the actual world ('artrite', in the example) could still count as the same word used in the actual world, otherwise Bert and TwinBert are just using different words, hence, given that the primary intension is related to the word somehow, the primary intensions differ. But in the possible world at which the two languages differ in this way, 'artrite' and 'arthritis' determine different extensions, and they could do so for a single bilingual speaker; therefore, they determine a different function from centered possible worlds to secondary intensions.

The two-dimensionalist might point out in reply that this problem would not prevent successful communication in the great majority of cases, due to sameness of secondary intension in the actual world. But, even granting the point about communication for the sake of the argument, we are still left an unwelcome result, because it seems to us that speakers of different languages can use terms meaning *exactly* the same as 'arthritis', and to have exactly the same thought, when they think they have arthritis, despite the phonological and orthographic differences; but this is impossible if the meta-linguistic move is accepted. Nor is the point dependent on focusing on public or stable meanings instead than on speaker or utterance meaning. Intuitively, the content of a particular utterance of 'John has arthritis', made by a particular English speaker, and the content of a particular utterance of 'John ha l'atrite', made by a particular Italian speaker, *can* be exactly the same, in all respects; and the same goes for two particular thoughts.

As we noted above, if the treatment extends to large part of our language, the counterintuitive result is multiplied. Moreover, the same result would apply to speakers of the same language who use different spellings of the word (either because two different spellings are accepted, or due to incompetence). Using 'arthrytis' instead of 'arthritis' would mean having a different concept. And indeed concentrating on the written form of the word seems insufficient, because surely illiterate speakers can have the concept; if we concentrate on phonological forms rather than orthographical ones, we avoid some of the problems, but then differences in pronunciation or accent would yield different concepts, which seems even worse. Not only that; it becomes impossible for a person affected by deaf-mutism since birth to possess exactly the same concept we do. The general problem is, again, that it seems obvious that phonological and orthographical features of the way we express a concept are inessential to it.

B) A second point is internal to the theory. Chalmer's account of Bert's case seems to undermine at least a significant part of the motivation for two-dimensionalism. I take one main motivation for two-dimensionalism to be the possibility to ground a priori knowledge of conceptual and metaphysical possibilities; but there is very little Bert can know a priori about arthritis. By the specification of the primary intension given above, next to nothing about arthritis can be learned. Surely Bert, according to Chalmers, could not know a priori that arthritis cannot apply to his thigh. But if my reasoning is correct, if he could not we cannot either, because we would also have a deferential concept; and if most concepts are like that, there is very little we can know a priori in general. If we defer about 'arthritis', then it is not clear why we wouldn't defer about 'consciousness', which is a term having so much more

empirical and conceptual complexity connected with its use. But then the primary intension of 'consciousness' would be scarcely able to play any theoretical role. Of course Chalmers would argue that 'consciousness' is special in this respect. But even granting for the sake of the argument that this is so, it seems enough of a problem, from the point of view of the two-dimensionalist, that most of us are not in a position to know a priori most things we thought we could, e.g. that a bachelor is an unmarried man. It is easy to think of cases on the lines of the arthritis case for 'bachelor'. Supposing 'bachelor' necessarily applies to all and only unmarried men, we can easily imagine someone being convinced by a misguided dictionary that 'bachelor' does not apply to certain unmarried men, such as clergymen. Then we can imagine a twin lives in a community in which "bachelor" actually does not apply to clergyman. The same considerations that lead one to attribute a meta-linguistic concept to Bert and his twin, lead one to attribute a shared meta-linguistic concept here. And, as I argued above, the same considerations apply to normally competent speakers. The point about the vast number of cases we are dealing with is again relevant.

Perhaps one could try to alleviate the worry by thinking of a primary intension which is only partially meta-linguistic, such as 'the disease denoted by 'arthritis' in my linguistic community', a formulation which is suggested in passing in Chalmers [2007]. But this formulation already puts the concept in danger of being subject to a Twin-Earth case. Surely as soon as the primary intension is going to include interesting features of the phenomenon, this is going to become possible.

C) Finally, there is a risk that the kind of meta-linguistic primary intension we are discussing, as well as not allowing for a priori knowledge when this would seem desirable, allows some a priori knowledge that seems not desirable. If we are in a position, letting for simplicity the primary intension of 'water' be roughly expressed 'the waterish stuff of our acquaintance', to deduce a priori 'there is waterish stuff of my acquaintance in the glass' from 'there is water in the glass', then we should be in a position to deduce a priori 'the phenomenon denoted by 'arthritis' in my linguistic community occurs in my knees' from 'I have arthritis in my knees'. But this seems a most unwelcome result, because it also allows one to then deduce the existence of a linguistic community, and of the word 'arthritis'; but it is counter-intuitive, to say the least, that such an inference could be a priori justified. The objection is not that I gain a priori knowledge of a contingent proposition.[44] The point is just

[44] Schroeter [2005] also argues that Chalmers is committed to our having a priori knowledge of the existence of a linguistic community, on grounds not strictly related to 'deferential' concepts. She also correctly notes that Chalmers takes this to be a serious objection to different interpretations of the two-dimensionalist

that this particular inference, from 'I have arthritis in my thigh' to 'there is a linguistic community' is obviously not a priori justified; in fact, crucially, it is obviously not justified at all, and the same applies to the corresponding conditional, 'if I have arthritis in my thigh then there is a linguistic community'. Standard competent speakers judge the inference invalid, and the conditional false. It is probably useful to compare this problem for two-dimensionalism with a similar problem that has been pressed against externalist views, the McKinsey problem[45]; the externalist is committed to accepting as true, and a priori knowable, conditionals such as: 'If I think that I have arthritis in my knee, then there is a linguistic community'. Chalmers, as far as I can see, is certainly committed to accepting that as a priori. But I am not pressing whatever objection derives from this, since I am assuming that no decisive objection can be canvassed against externalism on that basis, even it were established that externalists are committed to the apriority of the conditional. The objection I am pressing is that Chalmers is also committed, unlike the externalist, to accept as true, and a priori knowable, the conditional 'if I have arthritis in my knee, then there is a linguistic community'; and the latter conditional seems much worse than the former, in that it seems the antecedent and the consequent are completely unrelated.

A reply to this point is hinted to in Chalmers [2002], where he writes:

> If Bert uses his semantically deferential concept to think *my father has arthritis in his thigh*, how can we evaluate this thought in a world in which there is no word 'arthritis' in Bert's language, and in which Bert has no thoughts about his father's health? On the epistemic framework, it is most natural to say that the epistemic intension of Bert's *arthritis* concept picks out nothing in this world. In effect, the use of a semantically deferential concept *presupposes* that one lives in a community that uses the relevant term, just as a notion such as *The present king of France* presupposes that there is a king of France. If I discover that those assumptions do not hold in my actual scenario, it is reasonable to judge that my thoughts involving these concepts lack truth-value. The same goes for alternative scenarios. In Bert's case, the epistemic intension of his thought is indeterminate in the relevant worlds. In effect, Bert's thought partitions the space of scenarios in which the background assumptions are satisfied, and says nothing about those worlds in which the assumptions are false. (Chalmers, [2002], p. 626)

---

framework. However, given that the framework in general allows for contingent a priori knowledge, Chalmers could accept this as one such case, and insist that the advantages of his version of two-dimensionalism lie elsewhere. Aprioricity is after all a theoretical notion. But the truth of the conditional is instead a consequence which clearly flouts our pre-theoretic judgements. This way of presenting the objection parallels Kaplan's objection to token-reflexive accounts of the semantics of indexicals and demonstratives. See Kaplan [1989] for the original argument, and Predelli [2006] for developments of that debate.

[45] See fn. 39 for references.

However, the model of presupposition failure does not seem at all to fit this case. First note that the point remains that if 'Bert's father has arthritis in his thigh' was true then, according to this view, 'there is a linguistic community' would also have to be true, for if the latter was false the former would lack truth-value, and hence would not be true. In fact, this consequence is entirely acceptable in the cases which provide most plausible examples of semantic presupposition; the inference from 'the present king of France is bald' to 'there is a king of France', and the corresponding conditional, seem completely unproblematic, or at least vastly more plausible than the corresponding inference and conditional which Chalmers is committed to.

Finally, note that objection C would not be automatically answered even if we allowed the primary intension to contain not only meta-linguistic material; provided that satisfying the meta-linguistic criterion is a necessary condition for the term to apply (leaving the extension empty or undetermined where the condition is not met), the inference should still be valid and a priori justified.

2.3.3. Uncomfortable Sofas

Before moving to consider the consequences of the rejection of 1, and some possible replies, there is another (classic) thought experiment it will be useful to consider. The thought experiment is the one described in Burge [1986]. It also involves someone having a seemingly a priori false belief, but, unlike in the arthritis case, there is no linguistic incompetence involved, or at least no deferential disposition. Instead the a priori false belief is grounded in sufficiently deviant background beliefs. We have seen that Chalmers' reply made some use of the fact that a deferential concept was involved in the previous case, so adding a counterexample in which that is not the case should reinforce the conclusion that 1 has to be abandoned. The case discussed here will be also relevant in section 2.4.1. Burge describes the case as follows:

> We begin by imagining a person *A* in our community who has a normal mastery of English. *A*'s early instruction in the use of 'sofa' is mostly ostensive, though he picks up the normal truisms. A can use the term reliably. At some point, however, *A* doubts the truisms and hypothesizes that sofas function not as furnishings to be sat on, but as works of art or religious artifacts. He believes that the usual remarks about the function of sofas conceal, or represent a delusion about, an entirely different practice. *A* admits that some sofas have been sat upon, but thinks that most sofas would collapse under any considerable weight and denies that sitting is what sofas are pre-eminently for. *A* may attack the veridicality of many of our memories of sofas being sat upon, on the grounds that the memories are a product of the delusion. (Burge [1986], p. 707)

Burge also invites us to imagine that *A* is going to look for confirmations of his hypothesis, in any rational way you can think of, and if he were to find none he would admit his error. We can add to the story, to convince us that *A* is not to be disregarded as a rational subject, that *A* has got misleading but reasonable evidence that something (perhaps a conspiracy, or an extra-terrestrial civilization, or an evil demon) is inducing the delusion upon human beings.

The further step of the thought experiment is imagining *B*, an individual internally identical to *A*, who lives in a possible world where indeed there are no sofas (not in the sense we standardly think of them), but only works of art or religious artefacts, looking superficially like sofas. By a series of coincidences, *B* is brought to falsely think that these objects are commonly sat upon, and they are meant to be sat on. In his community, most people know that the primary function of these objects is to be works of art or religious artefacts; but all *B* has heard is identical, at least in sound, to what *A* has heard. However, by the time *B*, having acquired new evidence (subjectively indistinguishable from the evidence *A* has acquired), is doubting that 'sofas' are meant to be sat upon, his doubts are correct. What he expresses with that sentence is something false.

The conclusion that we should draw from this thought experiment, for our present purposes, is the following: *A*'s belief is a priori false (remember that the idea of an a priori false belief, *per se*, is entirely unproblematic in the two-dimensionalist framework), while *B*'s is true. In a two-dimensionalist framework, this entails the primary intensions of their beliefs differ, despite *A* and *B* being internally identical. The difference with the arthritis case is that *A* is not plausibly regarded as linguistically incompetent. He is informed of the prevailing pattern of use of 'sofas', and he knows dictionary definitions (although he thinks they are mistaken). He is capable to use 'sofa' in the standard way for most purposes, and to engage in sophisticated conversation about the nature of sofas, in which he and his opponents do not seem to talk past each other; his error seems motivated not by misunderstanding but rather by his non-standard background beliefs. Most importantly, he is not disposed to defer to anyone in his use of 'sofa'.

What could the two-dimensionalist reply at this point? It seems she has two possible strategies only: insisting again that *A* and *B* have a metalinguistic concept of sofa, or denying that *A*'s belief is a priori false. Let's consider them in turn. I'll be brief on the first strategy, because we found the metalinguistic move unsatisfying before, and here it seems to lose its main motivation, which was that the speaker was disposed to defer to experts about his use. *A fortiori*, we would have to find the move unsatisfying here.

The other response, claiming that *A*'s belief is false, but not a priori so, might come in two forms. The first form of this reply involves claiming that in general the belief expressed by competent speakers with utterances of 'sofas are meant to be sat upon' (the proposition expressed by the sentence in the public language, if you wish) is not a priori true. However, the example seems to make the move scarcely plausible. The truisms which are doubted could be varied to include, instead of a specification of a function of the artefact, facts about its typical perceptual appearances, or any other so-called 'prototypical' feature of the object. And the reply would incur at least one the problems discussed for the meta-linguistic strategy, that of making very little a priori knowledge available to speakers in general.

The other version of this kind of reply involves claiming that *A*'s utterance deviates significantly in meaning from similar utterances made by most competent speakers, so that his utterance in particular is not a priori false. This strategy will be considered in the next section. So far, the conclusion to be drawn is that we found no plausible way for the two-dimensionalist to save 1. If there are primary intensions, it seems they do not supervene on the internal states of the subjects entertaining them. We have not yet said what the consequences are for thesis 2, if thesis 1 is discarded. This will be the first topic of the next section. We will then consider whether it is plausible to defend 1 by appealing to 2.

2.4. Saving the Ideal (-ized Rational Reflection)?

I will consider here the possibility for the two-dimensionalist to defend some modified versions of the view, while accepting the counterexamples to 1 (sec. 2.4.1). I will then turn to the possibility of using 2 to defend 1 (sec. 2.4.2). Although I think that the move from the negation of 1 to the negation of 2 is fairly straightforward if we make certain other assumptions, that are held by the two-dimensionalist and perhaps by the defenders of conceptual analysis in general, I believe we should also consider the space of logical options for someone willing to stick to (some version of) 2, as it were, at any cost, since 2 is what is doing the epistemic work.

2.4.1. Partial Understanding

Although Chalmers [2002] explicitly advocates the internalistic nature of primary intensions, and this seems clearly the spirit of two-dimensionalism in general, maybe one could concede the falsity of 1 and still maintain a similar enough view to deserve the name of 'two-dimensionalism'. What would be needed is a strategy to maintain 2, while giving up 1. Let us remind ourselves what 2 says:

2. There is a primary level of content which supervenes on judgements about possible cases under idealized rational reflection (i.e. putting aside confusion, lack of time, and so on).

Is 2 incompatible with the negation of 1? If the relevant judgements about possible cases supervene on the internal state of the individual, then, since supervenience is transitive, the content that supervenes on those judgments will also supervene on the internal state of the individual, and so 2 will entail 1, and obviously the negation of 1 will entail the negation of 2. Why should we think though that judgments about possible cases supervene on the internal states of the individual? Consider Bert and TwinBert; on Idealized Rational Reflection (IRR), it seems, they would assent to 'I have arthritis in my thigh' at exactly the same centered possible worlds; therefore, 2 predicts that utterances of that sentence have the same primary content for them. It seems these considerations will generalize to any two internally identical subjects and any sentence; so 2 seems to entail 1. Obviously then the negation of 1 would entail the negation of 2. One might think that the two-dimensionalist could individuate judgments under IRR in a different way, so that they fail to supervene on internal states, and so Bert and TwinBert do not accept the same judgements. But what would be the motivation for accepting 2? There is a trivial way in which one could defend 2. If we individuate the judgments Bert makes by their content, the content will supervene on the judgments; not only judgments under IRR, but also actual judgments. I cannot assent to a proposition I am not in a position to entertain. It is not this triviality that the two-dimensionalist intended to defend. The idea was that judgments under IRR determine the set of centered possible worlds at which an utterance is true, simply because such judgments are correct, as it were, by definition. By stipulating enough into the idea of IRR, the result of saving 2 may be achieved, but its interest is limited. Clearly, an omniscient subject would judge correctly on any possible application of any concept, and so the content of the concept is modally tied to the judgments of the omniscient subject; but it is equally clear that there is no explanatory value in this observation. We will come back to the problem of understanding IRR in the next section.

A further possibility, closer to the spirit of epistemic two-dimensionalism if not to its letter, would be to modify the framework by re-introducing the distinction, which we considered in the section 2.1 above, between partial and full understanding, and limiting both 1 and 2 to subjects with full understanding. Some of the problems that I discussed in section 1 will re-emerge; however, the two-dimensionalist is independently committed to saying that deviant experts, like McGee, would on IRR change their mind, or else they are using a

different concept altogether. I will come back to the plausibility of this strategy in the next section. But at least the distinction between partial and full understanding could be thought to have a point about the arthritis case, since there seems to be some pre-theoretical plausibility to the claim that Bert only has partial understanding of the term. Let us see how this strategy would work. Consider first Bert and TwinBert; TwinBert, it seems, on IRR will arrive at the conclusion that 'arthritis', as he uses it, could possibly apply to the ailment in his thigh; and that conclusion will be correct. How about Bert? It might be claimed he is not in a position to start, so to speak, IRR, because by hypothesis he has only partial understanding of his own concept. If he were to have full competence, he would arrive correctly at the conclusion that 'arthritis' could not possibly apply to the ailment in his thigh. But then having full understanding of a concept will not be itself a condition supervening on the internal states of a subject. Or else we need to deny that TwinBert has full understanding; but then, by analogy, for the considerations discussed in section 2.3.2 above, full understanding will be very rare indeed. The more general problem is that we don't seem to have a grasp of 'full understanding' which is not just 'understanding plus the avoidance of error'. All these questions would need to be addressed if one wanted to make use of the distinction between partial and complete understanding in the context of a two-dimensionalist theory of content.

Even if one were ready to introduce the distinction and able to clarify it in a useful way so that it could apply to Bert's case, there would still be a major problem for extending this strategy to *A*'s case. Like Bert, it would have to be claimed, *A* has an imperfect understanding of the concept and therefore is not in a position to engage in IRR. However, *A* does not need linguistic information to fill the gap between partial and complete understanding, he needs factual information, which is plausibly regarded as empirical. With the needed information, *A* would come on IRR to the correct conclusion. Now, though, if this implies that judgements about possible cases on IRR are not a priori, this consequence seems to undermine one main motivation for two-dimensionalism, which we mentioned before, i.e. that it is supposed to ground a priori knowledge of metaphysical possibilities. But at this point, given that the notion of primary content in the sense of 1 has already been abandoned, very little interest would seem to remain in epistemic two-dimensionalism and its motivations.

Could the two-dimensionalists deny that, on this reply, judgements under IRR lose their a priori status? One way they could try to do that is by invoking the distinction between the enabling and the evidential role of experience in grounding a priori knowledge. Experience, it is commonly held, can play the enabling role, i.e. permitting the subject to

acquire the relevant concepts, without playing any evidential role in supporting a claim, and hence without depriving the claim of a priori justification. In chapter 3, I will discuss the shortcomings of this distinction, expanding on a challenge already put forward in Williamson [2007]. But here the only claim I need is that the enabling/evidential distinction breaks down if one combines it in a particular way with a partial/full understanding distinction. On the imagined view, one can have partial understanding of a claim, without having the grounds to judge it to be true, which would be acquired only with full understanding. The subject needs, in other words, empirical information to move from a false belief to a true one. But then the resulting true belief would be paradigmatically a posteriori.

2.4.2. Internalism Strikes Back, in a Circle

If one is convinced that 2, which is central to two-dimensionalism, is untenable without 1, and one wants to save two-dimensionalism, one might perhaps think we have been too quick to give up 1; we could have explored a different strategy, i.e. using 2 to defend 1.

The move of re-interpreting Bert's and *A*'s words as expressing a non-standard concept is not unmotivated after all, it could be claimed. It is motivated exactly by the fact that they would express non-standard judgements on IRR. In particular, they have the same concepts TwinBert and *B* have, since they share judgements under IRR. In the case of Bert, this would involve claiming the belief he expresses saying 'I have arthritis in my thigh' is true after all. In the case of *A*, presumably the primary intension of his concept would be something like 'the objects of sofish shape of my acquaintance'; this would pick up sofas in our world, and so, if he forms the belief that 'sofas are not meant to be sat upon', that belief would be false. However, it would not be a priori false, while we are assuming that the content standardly expressed by competent speakers with the same (syntactically individuated) words could be known a priori to be false.

The problem with this move is that it risks being a sort of stipulation. Of course it is perfectly legitimate for the two-dimensionalist to define primary intension in such a way to make it supervene on judgements under IRR. Then we also note that judgements under IRR plausibly supervene on internal states. The conclusion we should reach at this point is that primary intensions, *if they exist*, supervene on internal states of the subject. However, if there are cases in which two subjects intuitively differ in the content their concepts have, while their judgements under IRR would be the same, or, the other way round, their contents are intuitively the same, while their judgements under IRR would plausibly diverge, it is not a good reply that such cases are impossible by definition. The sort of content the two-

dimensionalist has defined might after all fail to have any psychological reality, or semantic role. Two-dimensionalism took its initial motivation from its ability to account for the (alleged) pre-theoretic judgement (or intuition) that in TwinEarth-like cases there is a level of content in common between the two subjects, which externalism cannot capture. It would therefore seem extremely awkward if in order to apply the two-dimensionalist theoretical machinery we had to deny, in a large number of other cases, our pre-theoretic judgements of sameness or difference in content. But this seems to be clearly the case at least for the envisaged re-interpretation of Bert's beliefs; to claim he has a true belief about the disease he has at his thigh is just to ignore our pre-theoretic judgements, as it were. He has a false one, thinking it is arthritis. Moreover, to have the true belief he would need the concept TWARTHRITIS that is used in TwinBert's community. And this applies to *A* as well; in order for his beliefs to b contingently false, he has to use *B*'s concept. But how did they acquire these new concepts? Surely not by being taught. Nor did they ever intend in the slightest to create a new concept. They simply should have acquired the new concept by going wrong in the application of the old one; but this is quite incredible in these specific cases. Surely we can be inclined to say about *a community* of speakers that goes wrong *for a long time* and *in a systematic way* about the application of a concept that they created a new one; a new norm of application has taken over the old one. But how could Bert's or *A*'s peripheral mistakes have such an effect? And what is the advantage of attributing to them such new concepts? These questions seem to have no answer, unless the stipulation that primary intensions supervene on IRR is invoked. But appealing to that stipulation would beg the question here, since primary intensions, defined that way, might not figure at all in the best explanation of Bert's or *A*'s mental states and linguistic behaviour.

In addition to the imaginary cases discussed so far, consider the following real case of theoretical disagreement[46]: Most epistemologists hold that knowledge entails belief; it is not possible for a subject to know that *P* if the subject does not believe that *P*. However, A minority of epistemologists deny this claim. It is surely implausible to charge these people with linguistic deficiency about 'knowledge' or 'belief'. There is clearly a disagreement here; the theorists defending the minority view therefore do not have a deviant concept, otherwise there could be no disagreement, but just talking past each other. Again, to stipulate that in this kind of cases the deviant theorist must have created a new concept is unmotivated. But the two-dimensionalist does not need to say that; she might claim that on IRR the theorists who

[46] The example of the disagreement between McGee and defenders of *modus ponens* considered before would serve equally well.

are wrong, let us suppose they are the ones denying the entailment between knowledge and belief, would recognize their error. If this is so, the disagreement is real, because they are wrong on the primary intension of their own concept, for lack of adequate reflection. The case would be analogous to one of two mathematicians disagreeing about the truth of some complex mathematical statement.

However, it's not clear what guarantees that in typical cases of theoretical disagreement the theorists who are going wrong would change their minds on IRR any more than the subject in the sofa case. The evidence that we have so far indicates that they believe what they believe stably, on reflection; e.g. they believe it is not a priori impossible for a subject to know something she does not believe, and in fact they believe, on reflection, such a possibility to be actual. Of course, we might insist that, as in the mathematical case, the theorist would, given unlimited cognitive resources, come to the correct conclusion. At this point, though, I think we have lost our grip on the notion of IRR. Clearly, we do not have access to this notion except through the notion of what is a priori possible for a subject[47]. What is a priori possible for a subject should be determined by the primary intensions of her concepts; and the primary intensions are determined by what is possible under IRR. We are moving, it seems, in a circle of definitions; this might not be itself a problem, but it becomes a problem if we are supposed to motivate the counter-intuitive claim that there is no disagreement in some cases (the arthritis case, the sofa case, the latter case of theoretical disagreement) by appeal to what conclusion the subject would reach on IRR. There are plausible cases of disagreement about what is possible, e.g. about the possibility of knowledge without belief, which would not plausibly disappear on IRR. To stipulate this situation is impossible for some kind of content only makes it implausible that this is the kind of content that our concepts and words possess. To reiterate, two-dimensionalism gets some initial plausibility from its alleged ability to do a better job than other theories in capturing judgements about content. If it ends up forcing our judgements to be discounted in favour of some general principle, which is required by the two-dimensionalist's theoretical commitments, then there is little to be said in its favour.

[47] Chalmers is explicit about that; see e.g. Chalmers [2006b] section 3.3, and [2006a] sections 3.3. and 3.9; in the latter work, Chalmers makes the circle smaller by mostly avoiding talk of IRR and connecting primary intentions and apriority directly.

Conclusion

We began this chapter by describing a certain picture of conceptual analysis, on which our judgements on philosophical thought experiments teach us about our concepts, in virtue of their being (in some sense) epistemically analytic. However, we saw that there are good reasons to think there are not, or not nearly enough, epistemically analytic truths to make that picture attractive. I then considered epistemic two-dimensionalism as a strategy to overcome such difficulties. Two-dimensionalism also incorporates the view that there is a fundamental level of content supervening on the subject's internal states. However, I have considered some thought experiments, mostly the ones presented in Burge [1979] and [1986], which are largely neglected in the literature on two-dimensionalism, and have found the two-dimensionalist defense of the idea that there is such a kind of content unsatisfying in the light of these thought experiments. In particular, I considered the response to the thought experiments provided in Chalmers [2002], and I found it implausible on independent grounds. I have then considered the possibility of conceding that point and still holding to a modified version of two-dimensionalism, involving a distinction between partial and full understanding. However, it seems this would require deep modifications of the view, and the resulting theory seems unstable, and unmotivated. Finally, I considered the possibility of going back to defending primary content by appealing to its definition in terms of judgements under idealized rational reflection, but I've found no reasons to think that content of such a kind is what human beings use in thinking or talking.

My conclusion so far is that the idea of conceptual analysis as I described it at the outset is also to be abandoned. Of course, this does not mean that we should stop considering imaginary cases. On the contrary, my arguments here rested on consideration of some such case, and I will say something positive about the epistemology of our judgements on thought experiments in chapter 4. Before doing that, however, I will discuss in the next chapter the consequences of the conclusion of this and the previous chapter on the notion of apriority. One might still think that, although not based on intuition, and not epistemically analytic, our judgments on philosophical thought experiments are still 'exceptional' in that they are a priori, in the sense, roughly of being independent of experience. I will argue in the next chapter that there is no useful way to specify the notion of independence from experience, thereby concluding my case against philosophical exceptionalism.

Chapter 3
A Priori Knowledge: The Last Dogma of Rationalism?

Experience is not what happens to a man. It is what a man does with what happens to him.

Aldous Huxley

Introduction

In the previous chapters, I have argued that there is no epistemic role played in philosophy by 'intuitions' or seemings, and that philosophy does not involve something like conceptual analysis, in the sense of the discovery through thought experiments of a certain class of judgments which are linked psychologically or epistemically to the possession of the concepts involved. One might agree with the conclusions reached so far, but still claim that philosophy has one significant epistemic feature that separates it, together with disciplines such as logic and mathematics, from the empirical sciences, the feature of producing a priori knowledge. I will argue here that the traditional distinction between a priori and a posteriori has not been, and most likely cannot be, spelled out in such a way to ground this distinction. This is not, of course, to deny that there are differences of various sorts between philosophy and the empirical sciences. It is instead to deny, first, that the differences can be captured in an illuminating way by the distinction between a priori and a posteriori, and, second, that the differences have a clearer epistemological significance than the similarities. The more general moral of the chapter will be that there are serious reasons to doubt that the a priori/a posteriori distinction has any epistemological use.

In a way, the problem with the a priori/a posteriori distinction stems from the conclusions of the previous chapters. Epistemic analyticity and intuitions have both been considered sources of a priori knowledge. However, the rejection of the a priori/a posteriori distinction does not trivially follow from the conclusions reached in previous chapters. Someone might think that a priori knowledge is untouched so far; for example, take the

following recent description of the a priori by Tyler Burge, which I find at the same time sufficiently representative of the prevalent idea of apriority in the philosophical community, and sufficiently careful:

> For a claim or belief to be warranted apriori is for the warrant not to depend for its force on sense perception, or other sensory material, or on perceptual belief. The force of the warrant normally rests instead on understanding or reason. The explanation of the warranting support does not appeal even partly to sense perception, sensory material, or perceptual belief. It appeals to understanding or reason. (…)
>
> Claims of apriority are commonly misunderstood. Misunderstanding is a primary source of resistance. To say that a belief is apriori is not to say that it is innate, obvious, rationally certain, indubitable, infallible, imposed on the world, or immune to revision (even empirical revision). To say that a belief is apriori is not to say that it is easy to recognize that it is true, or that it is uncontroversial—much less that it is easy to recognize that it is apriori. (Burge [2010] p. 534)

Burge does not appeal to intuitions or analyticity, neither explicitly nor implicitly, in any of the senses discussed in the previous chapter. A priori warrant is characterized negatively by independence from experience, and positively by involving understanding and reason. Note that no link is made between understanding and beliefs or judgments, but only between understanding (and reason) and the warrant that certain judgements have.

Unfortunately Burge does not say much about how we can further understand at least one of the crucial notions, that of a warrant *depending* (or failing to depend) on sense perception. I will discuss various ways of spelling out that notion. The overall argumentative strategy of the chapter (largely inspired, as will be made clear, by Williamson [2007] and Hawthorne [2007]) will be the following: given that there is no epistemic analyticity, there is no principled way of separating the epistemically relevant components of the belief-formation process from the non-epistemically relevant ones such that it provides a distinction extensionally similar to the traditional one. In particular, no way of specifying the notion of independence of experience allows what is traditionally thought to be a priori and a posteriori to have a justification, respectively, independent of, and dependent on, experience. In some more detail, the plan of the chapter is as follows. In section 3.1, after having explained the terminology I am using, I will distinguish the problem I will be discussing from various other issues related to the notion of a priori, and, most importantly, lay out the main problem for the a priori/a posteriori distinction by illustrating various problematic cases. The problematic cases are mostly pieces of knowledge which were traditionally considered a posteriori, but end up being a priori on a certain standard account of independence from experience. In section 3.2, I will consider a proposal by Carrie Jenkins, that of linking the relevant notion of

dependence to the notion of epistemic grounds. I will show that this does not improve on previous suggestions. In section 3.3, I will discuss the idea of positing restrictions on the modal status of propositions that can be a priori known. I will argue that the proposal is problematic in various ways, and ultimately unmotivated. Finally, in section 3.4, I will consider the proposal that, in cases of a priori knowledge, understanding the proposition, although it is not sufficient for justified belief, provides a basis for justified belief, in the following sense: given that one understands the proposition, one needs only to engage in some reasoning to know that proposition. Firstly, I will argue that such reasoning should be itself not dependent on experience. After that, I will argue that substantially the same problems that emerged in spelling out the independence of experience requirement imposed on the justification of beliefs will resurface if we try to spell out an independence of experience requirement for reasoning. I will try to illustrate the point in particular using an example of philosophical knowledge.

Before going into the issues, I need to say a few words, although I will not be able to do justice to the historical and theoretical issues involved, about the relationship between the challenge to the a priori/a posteriori distinction I am going to talk about in this chapter and the similar challenge that has a (more direct) Quinean origin. For one might worry that, after all, the distinction between a priori and a posteriori was successfully rejected a long time ago, and therefore we do not need to spend more time on it. I believe, as the previous chapter made clear, that the best motivations for the rejection of epistemic analyticity are strongly connected to the motivations for accepting externalism about content, and I would say the same about the motivations we have to reject the a priori/a posteriori distinction. Quine's motivations for the rejection of the analytic/synthetic distinction (and of the a priori/a posteriori distinction, which he assumed, as most philosophers of his times, could not be spelled out independently) are, at least superficially, quite different. This is not to say that Quine's arguments were bad. They were given in a different historical context, with different background assumptions. They have been deservedly influential, but they certainly did not command universal assent, and much less do they do so nowadays. I believe the arguments given here are, in the end, in the spirit of Quine's arguments, only they are different (and, I hope, even better) because I stand on the shoulder of several giants, among whom there is Quine's himself. As I said, I am not going to be able to give a real comparison. However, there is a difference that I would like to note, which has to do with the way the conclusion of my discussion will be phrased. Quine, and even more clearly some Quineans (see for example Devitt [2005]), often say that all our knowledge is empirical (where 'empirical' is just

synonymous of ‘a posteriori’). Now, it seems to me that this is a misleading way to put what they ultimately want to say. If we can affirm that all knowledge is empirical, then we are employing a sufficiently clear meaning of ‘empirical’, presumably on the line of ‘epistemically dependent on experience’. I think that in a way this is conceding the opponent, let us call her the ‘rationalist’, too much. I do not think we have been given by the rationalist a clear enough account of dependence on experience to judge whether some of our knowledge is so dependent. Moreover, and more importantly from the dialectical point of view, if the sense of ‘dependent on experience’ that the empiricist is using is such that all knowledge is dependent on experience, then probably the rationalist will, and should, claim to be using a more restricted sense. It is hard to deny that the rationalist can stipulate a sense of ‘dependent on experience’ in which some knowledge is not so dependent. After all, everyone agrees that there are differences in degrees between various parts of our knowledge, with respect to how closely they depend on experience. So what can stop the rationalist from finding a cutting point? Nothing, in principle, is my reply. But I am going to argue that there just is no cutting point that will satisfy the rationalist, there is no way of cutting that will leave on one side what has been traditionally thought a priori and on the other what has been traditionally thought a posteriori.

3.1. The Challenge

Most analytic philosophers, despite the presence of a contrary tradition coming mainly from Quine, make use of the distinction between a priori and a posteriori. Not only is this distinction customarily taught in introductions to epistemology, but it plays a crucial role in many areas, for example in debates about physicalism in philosophy of mind, and, not least, in debates over philosophical methodology. My focus here will be on arguments that, unlike Quine’s, attack the distinction on the background of externalism about content. My aim is to consider and refute some responses to this kind of arguments which concede semantic externalism, or at least do not directly attack it, but still try to save the distinction. In this section I will introduce some terminology, and then I will illustrate the challenge. The next three sections will consider three kinds of reply to the challenge, and I will argue that they all fail.

Being a priori is traditionally thought to be primarily a characteristic of knowledge, or justification, that of being independent of experience. A justification, or a piece of knowledge, will be a posteriori just in case it is not a priori. At some points, it will make a difference whether we think of justification or knowledge as the primary bearer of the property of being

a priori. If there is a viable notion of justification which is not itself derivative from the notion of knowledge, then we can define a priority primarily for justification. A subject has a priori justification for a belief if and only if the subject's justification is independent of experience. We may then say that a subject knows a priori that *p* if and only if the subject knows that *p* and the justification she has for her belief that *p* is a priori. A proposition can be said a priori, in a derivative sense, if it is possible for a subject to have an a priori justification for believing it. Note that it is compatible with these definitions that you have an a posteriori (not a priori) justification for an a priori proposition. If we wished to focus on knowledge from the start, we could just say that a subject knows a priori that *p* if and only if she knows that *p* independently of experience. At most points, anything I will say could be reformulated accordingly. However, I will mostly use the definition in terms of justification, which is slightly more common.

Let us assume for now this purely negative characterization of the a priori, as independence from experience (I will consider ways to modify it in sections 3.3 and 3.4). It has often been recognized that it is not simple to spell out exactly what counts as independence from experience. But there are several issues concerning the latter notion, and I believe the debate has often avoided the most fundamental problem. I will briefly describe three issues, of which the first two seem to have attracted most of the discussion so far[48], with the main purpose of distinguishing them from the topic of this chapter. The first issue is whether the independence should entail that no possible experience could undermine the subject's justification. The second is whether we should count as relevant features of the subject's past experience just whatever content the experience has, or we should also count as relevant the fact that the subject's experience has no further content (or, in other words, the absence of certain experiences). The third issue, of which theorists have always been aware (since Kant at least), but has been thought of as a problem only recently, has to do with the role of experience in acquiring the concepts which are involved in the content of the belief (the content which, if everything else is fine, is known). Most theorists who discuss the issue are happy to recognize that experience has to play a role in the subject's acquisition of the relevant concepts, but they think this role can be sharply separated from the role of experience which makes a justification empirical. This latter issue is connected to what we are going to discuss here. But the question we are discussing is, as I see it, more fundamental. The question is about the exact nature of the dependence relation we are talking about. From now on I will call the relation (possibly a disjunction of various other relations) in virtue of which

[48] See e.g. Casullo [2003], especially chapters 2 and 3, and Field [2005], and references there.

justification depends on experience, the Dependence Relation (DR). Understanding DR seems crucial to a good understanding of the a priori.

The difficulties involved in the latter aspect of the problem have been, I believe, often underestimated. The general challenge which I am going to present, drawing from the work of Hawthorne and Williamson, is the following: there is no characterization of DR such that it delivers results sufficiently close to what has been traditionally considered a priori, and a posteriori, to make it worth continuing to use the distinction. However, I will not give a deductive argument to the effect that such a characterization is impossible. We will just consider a number of prima facie plausible options, and I will argue that they fail. For any initially plausible characterization of DR that I know of, either many things which are traditionally counted as a priori count as a posteriori, or vice versa. Now, this of course does not show that the distinction, or, to be more precise, the various distinctions which result from different characterizations of DR, are incoherent, or even that one side of any such distinction is always going to be empty. But they will not give us anything close to what has been traditionally thought to be the extension of the concepts of priori and a posteriori justification. I have no argument against accepting one such revisionary notion, but I see no particular interest in it, and I think defenders of the a priori would not either. A more modest conclusion which I will aim to establish is the following: there is no characterization of the dependence relation such that it both makes the justification afforded by typical philosophical thought experiments count as a priori, and it does not count as a priori many things which were traditionally thought to be a posteriori.

Given the nature of the challenge, it will be necessary to have a fairly wide range of examples to look at. A proponent of the challenge needs show that the various ways of spelling out the distinction have either the consequence of classifying as independent of experience much knowledge that was traditionally considered a posteriori, or *viceversa*. Of course, human knowledge is (hopefully) quite vast and diversified, and we cannot literally look at a significant proportion of its whole body. What is crucial is that the examples we are going to look at are not atypical in some way, and that it is easy enough to see how the same reasoning would apply in a vast number of cases. In this section, I will start with a fairly standard, and perhaps superficial, characterization of the dependence relation, and look at various cases that are problematic for that characterization. The following sections will argue that more sophisticated characterizations do not fare better.

My initial characterization of the way defenders of the a priori/a posteriori distinction explain independence from experience is taken from the following passage by Bonjour:

> Negatively, an a priori reason for thinking that a claim is true is one whose rational force of cogency does not derive from experience, either directly (as in sense perception) or indirectly (by inference of any sort – deductive, inductive, or explanatory – whose premises derive acceptability from experience) (Bonjour [2005] pp. 98-99)

Now, to anticipate my conclusion, it will turn out that this characterization of dependence, on a standard reading of 'inference', will make a lot of what we thought a posteriori justified end up a priori justified. To illustrate what I have in mind when I refer to a standard reading of 'inference', consider, for example, the following dictionary definition (Merriam Webster Online Dictionary) of inference: "the act or process of inferring as a: the act of passing from one proposition, judgement or statement considered as true to another whose truth is believed to follow from that of the former b: the act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty."

To see the definition above has the consequence of making too much a priori, consider the following counterfactual (the example is from Williamson [2007]):

1) If two marks had been nine inches apart, they would have been further apart than the front and back legs of an ant.

Williamson comments: "One could know [1] without remembering any occasion on which one perceived an ant, and without having received any testimony about the size of ants. The ability to imagine accurately what an ant would look like next to two marks nine inches apart suffices."[49] In the situation envisaged in the comment, belief in 1 is arrived at through a process which is neither direct perception nor inference from direct perception, since it is not inference in any ordinary sense. It might still be thought that there is a sort of unconscious or implicit inference at work in this case. This kind of suggestion will be ruled out in the following cases.

Consider also the following sentence (discussed by Williamson in a different context, exploring in general the epistemology of counterfactuals):

2) If the bush had not been there, the rock would have ended up in the lake.

[49] Williamson [2007] p. 167

Consider a context in which 2 is true, and known to be so by a subject uttering it. It is at least unclear that Bonjour's characterization is not met by the subject's belief. The subject does not directly perceive that 2 is true. Nor is it clear that the subject needs to go, or have gone at any point, through reasoning based on premises. Still, one might object, there are some experiences that the subject has which seem to play a role in sustaining the belief in 1; the subject perceives where the bush and the lake are, and so on. So, maybe, these perceptions make a direct, albeit only partial, contribution to the knowledge of 2. We can put this aside by considering the following variant of 2, or something along its lines:

2*) If things stand as I think they do as to the present position of the lake, the bush and the rock, then if the bush had not been there, the rock would have ended up in the lake.

As Williamson puts it, we arrive at knowledge in this case through a constrained use of our imagination. He considers the proposal that what we are really doing is relying on premises and drawing an inference, but he raises a number of objections. The main ones are related to the fact that we would be plausibly relying on principles of folk physics. However, firstly, we do not need to have conscious access to them; not just in the sense that they are not occurrent beliefs, but also in the sense that if presented with explicit statement of the principle we would not recognize it as something we believe. It is therefore unclear that it makes sense to talk of inference in describing the relevant belief-forming process. Secondly, principles of folk-physics are plausibly false, and even if they are true, it is unclear that they are known or even justified. The latter point will not be a problem if those principles only serve the purpose of constraining our imagination in a way that is reliable given the circumstances, but it should be a problem, epistemically, if they were used as premises, explicitly or implicitly.

A further example was proposed in a similar vein, although in a different dialectical context[50], by Philip Kitcher (see Kitcher [2000] pp.78-80). Consider

[50] The context of Kitcher's discussion is rather different from the present one. Kitcher, in the paper I quote from, is engaged in defending the view that a notion of a priori warrant that gives up a certain sort of infallibility cannot play the same role of the traditional notion. While I am sympathetic to Kitcher's view, it seems nowadays the overwhelming majority of the theorist who defend an a priori/a posteriori distinction do not claim any sort of infallibility of a priori warrant. So I find it dialectically not very useful to insist that they should really add infallibility to the notion, especially since that would have the consequence of making a priori warrant extremely scarce. So, in a way, it is more convenient to discuss the new notion on its own right.

3) A cubical die made out of uniform material has a chance of 1/6 of showing 6.

Kitcher invites us to imagine someone who comes to know 3 by reasoning, in the way one might do when introduced to probability theory, somewhat as follows: "Well, one of the faces will show, and the situation is completely symmetrical. Thus there's no reason why one face should be more likely to show than another. Hence the probability that the uppermost face will show a 6 is 1/6" (p. 78). It is clear that one, in such a case, is not either directly apprehending 3 by perception, nor making an inference from premises that are directly apprehended by perception. However, one might think 3 is rather different from 1 and 2*; one could just admit that 3 is a priori. But then, Kitcher objects, we should allow arguments that appeal analogously to symmetries and invariances in other fields to produce a priori knowledge, and these arguments have been used to prove "claims about sex-ratios in populations, the invariance of laws in frames of reference, even the existence of particles that fill particular places in a mathematical scheme" (p. 80). In other words, we would open the door to a surprisingly large amount of a priori knowledge about the world. To make this a bit more vivid, consider

3*) When a creature C is generated by two other creatures A and B of different sex, there is a chance of approximately 1/2 that C will belong to the same sex of creature A.

Why would we need some additional information, beyond understanding, to assess the truth of 3*, but not that of 3? There is no principled answer to this question. Yet 3* is, even more clearly than 3, something that is not traditionally counted a priori.

There is one more example that I wish to consider before discussing ways out of the problem. The example is brought up in Hawthorne [2007]:

> Suppose a teacher instructs someone in the laws of nature when he is young and then he remembers those laws later. There is a store in the memory bank that calls forth the laws on demand, applies them to various possible cases, thereby extracting conditional predictions, derived generalizations and so on. (We can think of two versions of the case. On one version, preservative memory of the laws is accompanied by memory of the teacher giving instruction in the laws. On another version, the laws are stored away without accompanying episodic memories.) Let us suppose that the person's nomic beliefs, conditional predictions, and so on, are highly reliable. Just fill out the details of the case in such a way that at least many of the relevant beliefs are safe enough to count as knowledge. (I assume here, of course, that if the mechanism of belief production is safe enough, then the believer will know its deliverances even if he cannot produce any empirical reasons whatsoever for the laws.) (Hawthorne [2007] p. 210)

Hawthorne then notices that there are two ways of individuating the relevant belief-forming process; the first includes the experiences involved in the learning, the second does not (both ways are available, in principle, whether or not the subject remembers those experiences). Therefore, if we individuate the process in the first way the subject's knowledge will count as depending on experience, and if we individuate it in the second way it will not. But then he observes that the situation is similar in the following case:

> Consider a case of mathematical knowledge. Here again, there are a variety of candidate processes to use as a test for experience independence. Let us focus on two: (a) Math-Short, which begins with the retrieval of mathematical information from preservative memory and ends with the application of computational techniques to answer the problem at hand. (b) Math-Long, which begins with the training in the relevant techniques and provision of the relevant mathematical information at school, home or college, and ends with the application of computational techniques to answer the problem at hand. Using Math-Short as our benchmark, the relevant belief is experience-independent. Using Math-Long as our benchmark, it is not. (Hawthorne [2007] p. 212)

Hawthorne claims that there is no principled reason why we should count the derived knowledge as a priori in the latter but not in the former case[51]. He considers one such possible reason, which is that in the mathematics case whatever training is sufficient for us to grasp the relevant thoughts, is also sufficient for us to have a justification for the belief. Hawthorne objects that this is not plausible in the light of semantic externalism, since the conditions for understanding a claim might be, and usually will be, too thin. We have gone through this sort of considerations at length in chapter 2. At this point, I believe that it has been shown that there is probably no proposition such that the understanding of it is in itself sufficient for a subject to have justified belief or knowledge of it; or at least that there are so few that no interesting epistemological theory can take them as the paradigmatic cases of a priori. However, I will not assume that one cannot defend a view on which the understanding provides in some sense *a basis* for the justification. Sections 3.2 and 3.4 discuss views of this kind.

[51] This is, it might be objected, just an instance of the generality problem for reliabilism, which resurfaces for safety-based approaches as well. I agree that there is a clear connection, but in no way the connection undermines the force of the worry for the a priori/a posteriori distinction. The point can be put as follows: why should a satisfying solution to the generality problem individuate processes in a relevantly different way in the two cases at hand?

Let us now consider some responses to the challenge[52].

3.2. Methods and Grounds

In revising Bonjour's account of DR, we need a notion somewhat broader than inference. Let us now use the term 'grounding' for such a notion. We then have that

G. A justification is a priori just in case it does not depend on experience, either directly (as in sense perception) or indirectly (by grounding of any sort).

Inference will be of course a form of grounding, but not the only one. In the case of 1* and 2 the belief in the counterfactual will be typically grounded in some past experiences. We will call what the belief is grounded on, unsurprisingly, the subject's grounds for the belief. I will turn later to what exactly this is supposed to mean. First though I want to illustrate how Jenkins [2008] uses this notion against Hawthorne's argument. She writes, about the case I described:

> …the force of this argument is supposed to derive from the idea that there is no natural choice between Math-Short and Math-Long when we are trying to say which is 'the' method by which the subject arrived at her belief. So there's no real reason to regard it as a priori or a posteriori.
>
> But even granting Hawthorne the point about methods, there is no analogous temptation to think that training and instruction at school, home or college might be said to form part of the grounds people have for their a priori mathematical knowledge. Maybe we should in the end conclude that they do, but that needs substantial argument. The natural pretheoretic view is that these things are not part of one's grounds. (Jenkins [2008b] p. 447. Original italics)

The first, dialectical, problem with Jenkins' reply is that it does not say anything explicit about the other case which Hawthorne was comparing to this one. So one could say that the

[52] There is one possible line of response which I will not discuss, inspired by Tyler Burge's work on the epistemology of testimony and memory (Burge [1993a], [1997], [1998]). On (a reading of) Burge's view, the role of testimony and memory is purely preservative. What this means is that testimony and memory, when working correctly, and in absence of defeaters, will preserve a particular content and, crucially, the epistemic status of that content. This view is particularly controversial for testimony. To illustrate, if I learn from a reliable mathematician about a theorem she has proved, on such a view, I acquire a priori knowledge of the theorem, even if I have no idea of where to start proving it. Malmgren [2006] provide what I take to be decisive considerations against Burge's view in the case of testimony. Moreover, it is not clear that, even if Burge was right, this would help, in particular with the sort of examples discussed by Williamson.

reply does not address the problem properly; Hawthorne never claimed that there was no reasonable interpretation of what is going on in the mathematics case such that it ends up constituting a case of a priori, independent of experience, knowledge. He only claimed that there is no such interpretation under which the other case does not get the same verdict. Does Jenkins point provide a way of separating the two cases? I must confess a preliminary difficulty I have in answering this question. I find Jenkins' claim on our "pretheoretic view" of what is ground for what a bit surprising. I am not sure that I have (or even used to have) a clear pretheoretic view on this particular matter. In addition, I worry that our judgements on what grounds what depends on our judgement on what we (have been taught to) consider a priori, rather than the other way round.

Moreover, insofar as there is a pre-theoretic concept of grounds, I think it counts as grounds only things that people are ready to cite in defense of their beliefs. For example, if I were asked something like 'what are X's grounds for her belief that P?' I would count considerations on X's psychology irrelevant, and I would try to spell out the reasons and the evidence that X is ready to offer. One reverts to psychological explanations of X's holding her belief, in this sort of situation, only when X has no grounds at all. But this conception of grounds would count as a priori both beliefs in the mathematics case and in the laws of nature case, since we are stipulating there are no experiences the subjects would cite. Of course this notion of ground is not fruitful from the epistemological point of view; having good grounds in that sense is neither necessary nor sufficient for one's belief to be justified[53]. It is, after all, a pretheoretic notion, although I am aware that it is not the pretheoretic notion Jenkins has in mind. But I am unaware of a different pretheoretic notion.

Be that as it may, is it the case that on a better understanding of grounds, we are able to differentiate between the two cases described by Hawthorne? To show that, we need some explanation of the difference. We need an account, at least a rough one, of the grounding relation. Jenkins acknowledges that the notion, as Hawthorne has pointed out, has a certain degree of unclearness, but she replies: "That it is difficult to say exactly what grounds are does not mean we may assume that another notion can serve as a proxy for these purposes."[54] I would like to make clear that I do not disagree with the quote just reported. I would find it unreasonable to ask to be told what grounds *exactly* are, or to be given a counterexample free analysis in terms of necessary and sufficient conditions. To be told nothing at all, on the other hand, would be unsatisfying. In particular, we need to have a sufficient grasp of the notion to

[53] Recognition of that goes back at least to Harman's insightful discussion (Harman [1973] pp. 24-33).

[54] Jenkins [2008] p. 445.

be convinced that it delivers the promised result, that of separating the problematic cases in a way that at least approximates the traditional one. But the problem is not so much that we do not have any account of the grounding relation; the problem is that nothing in the vicinity of the accounts which have been proposed so far will allow the notion to do the work that Jenkins is supposing it does. To show that, in the rest of the chapter I will discuss three broad ways of spelling out the grounding relation: the first is the causal-counterfactual theory, the second is the doxastic theory, and the third is some combination of the first two.

On the causal-counterfactual theory, a mental state x is part of the grounds for a belief if the belief is caused by x, or some counterfactual relation holds, such that if x did not obtain (or had not obtained), then the subject would not hold the belief. Typically, x will be a belief, but we can allow, in principle, any kind of mental state to enter the grounds for a belief. This rough description does not do justice to the subtlety of the various ways in which this sort of theory has been spelled out[55]; but those subtleties will not make a difference for our present purposes. The fact is that the theory will plausibly classify all our cases as not a priori, including those that were not thought to be so, since the learning experience seem to play a causal role, for example, in both Hawthorne's cases. However, the point is not so much that since learning is causally relevant it gets to be part of the grounds; there might be sophisticated versions of the causal-counterfactual account that avoid this result. The aim of such sophisticated versions is mainly to avoid deviant causal chains; but there is nothing deviant about the causal chains we are concerned with, or at least nothing more deviant in the case of knowledge of physical facts than in the case of knowledge of mathematical facts. The general point is that no reason has been given to think that the causal role of the relevant experiences has different epistemic effects in the two cases.

Similarly, consider these examples from Williamson, and compare them to 1, 2* and 3:

4) It is necessary that whoever knows something believes it.

5) If Mary knew that it was raining, she would believe that it was raining.

6) Whoever knew something believed it

Let us recall 1, 2* and 3:

[55] See Korcz [1997] and [2006] for a survey of the literature. I am here assuming that the considerations which apply to the so-called "epistemic basing relation" apply to the grounding relation. In absence of that, I do not know of any available account of the grounding relation.

1) If two marks had been nine inches apart, they would have been further apart than the front and back legs of an ant.

2*) If things stand as I think they do as to the present position of the lake, the bush and the rock, then if the bush had not been there, the rock would have ended up in the lake

3) A cubical die made out of uniform material has a chance of 1/6 of showing 6.

It is not clear that experiences and empirically justified beliefs do not causally sustain beliefs in these cases in the same sense that they do for 1, 2* and 3 when the latter are held in the sort of way that I described, a way which does not include conscious access to any sustaining belief.

The second broad kind of account of the grounding relation is the doxastic kind. On this sort of account, the subject must be aware of the supporting relation between her grounds and the relevant belief. This is not equivalent to saying that the belief has to be based on an inference from the grounds. For one thing, we are now allowing that the grounds are non-propositional. And even if the grounds consist of further beliefs, being aware of the connection seems a more liberal notion than drawing an inference. Since inference turned out to be a too narrow notion, it seems that this could help. But the notion is still too narrow, for our present purposes at least, leaving experience out of the subject's grounds in all the cases I discussed. For I take it that awareness of the supporting relation entails awareness of the relata. But, in all cases we discussed the subject may just not be aware of the relevant experiences, or of the relevant beliefs based on experience, and of their relation to the relevant piece of knowledge. So, while the causal-counterfactual view of the grounding relation plausibly counted too much as a posteriori, the doxastic view will count too much as a priori.

As far as I can see, mixed views will not improve the results outlined so far. For there are two natural ways one could construe a mixed view. One is by saying that the grounding relation holds just in case a conjunction of causal-counterfactual and doxastic requirements holds (as in Audi [1989]). Alternatively, one might say that the grounding relation holds just in case the disjunction of a causal-counterfactual requirement and a doxastic requirement holds (as in Korcz [2000]). But the conjunction of the two requirements, having the doxastic requirement as a necessary condition, has the same problem as the doxastic view, that of being too narrow in the way it individuates the grounds. On this view, anything we are not aware of falls outside the grounds, and therefore too much counts as a priori. On the other hand the disjunction of the two, having the causal-counterfactual requirement as a sufficient

condition, has the same problem as the causal-counterfactual view, that of being too broad in the way it individuates the grounds, and thereby counting too little as a priori.

It is sometimes suggested to me that what is driving the judgement that in the mathematics case we have a priori justification is that the subject *could* have worked out her mathematical knowledge by herself, so the knowledge does not depend *essentially* on the teaching. However, first it is unclear that this is true, and second it is unclear why it is relevant. It is unclear that it is true because, as Kitcher [2000] has argued, our mathematical knowledge might be essentially tradition-dependent, built up, so to speak, on the shoulders of the giants who developed the concepts and the techniques which shape our practice; so that there might be different traditions which lead us to the same true beliefs without such beliefs always constituting knowledge. It is unclear that the observation is relevant anyway, because what we are judging on is the epistemic status of the belief the subject actually has, not the epistemic status of the belief the subject might have had if she just reconstructed the concepts and the proofs constituting her knowledge by herself, assuming this to be possible. It might be that there is some sense of propositional (as opposed to doxastic[56]) justification in which all mathematical and logical truths have propositional justification by default, and therefore, in some sense, independently of experience. But this is not interesting if the beliefs we form on the basis of that propositional justification are never, or almost never, justified independently of experience.

Of course there might be a way of construing the grounding relation which I have not considered, and which is both independently plausible and adequate for drawing the a priori/a posteriori distinction. But then it needs to be spelled out, at least in rough outline. At present, the prospects for this kind of reply seem rather bleak.

3.3. Modal Content

A very different reaction to the Hawthorne/Williamson challenge is to just deny that in any of the examples they consider the relevant proposition is even a candidate for a priori knowledge, for a priori knowledge is limited to necessary truths, and the propositions involved there are all contingent (this is not uncontroversial in the laws of nature case, but we

[56] In the sense of Firth [1978] (but see also Harman [1973] pp. 24-33). Roughly, doxastic justification is a property of beliefs, the property of being formed in an epistemically appropriate way. Propositional justification is a property of propositions relative to an evidential state, that of being supported by the evidential state.

can certainly concede it for the sake of the argument). To consider this reaction at this point might look like a digression. But, aside from the intrinsic interest of the proposal, its relevance will become clear later.

Goldman [1999] considers some objections to the view that all a priori known propositions are necessary. Firstly, he considers the worry that some, following Kripke, want to allow for a priori knowledge of contingent propositions[57]; but he puts this worry aside, for the sake of the argument, since that view is controversial. I think this is too concessive, as I will argue later about a related proposal. Secondly, he considers the worry that most theorists do not want to make a priori justification factive; you can be subtly misled in the application of mathematical reasoning, and hence have a priori justification for a false proposition, and hence for a proposition which is not necessarily true. A possible reaction, considered by Goldman, is to weaken the relevant requirement for a belief to be a priori justified as follows: the subject holding the belief needs to have the further belief that the proposition is necessary. To this he objects that a subject could lack modal concepts altogether, and still have a justification for a mathematical proposition which we want to count as a priori. One could try to get around this problem by saying that the subject would believe the proposition to be necessary if she considered the question (this suggestion is inspired by an analogous move made by Pust [2000] about intuition[58]). But this fails too, for a subject could mistakenly believe that mathematical propositions are not necessary, due a deviant theory about the nature of mathematical objects[59], and it would seem entirely arbitrary to deny the subject has a priori knowledge for that reason. Similarly, Williamson thinks that a subject in a Gettier case does not know, but he does not think that necessarily a subject in a Gettier case does not know; however, it would seem unacceptable for the defender of the a priori to concede that Williamson, unlike other people, knows a posteriori that the subject in a Gettier case does not know.

A better suggestion that can be extracted from Goldman's discussion is the following:

Modal A Priori (MAP): A subject has a priori justification for a proposition *p* iff i) the subject has a justification for *p* independent of experience and ii) *p* is either necessarily true or necessarily false

[57] See Kripke [1980], Evans [1979], Hawthorne [2002].

[58] See Pust [2000] pp. 38-9.

[59] See Rosen [2002] for a detailed and convincing description of an entire (imaginary) community meeting such description.

If we use a notion of dependence on experience narrow enough to exclude the controversial cases, this definition does classify the cases we considered so far in the traditional way. For example, we could use Bonjour's account of DR described in section 3.1, on which a justified belief is a priori justified unless directly based on perception or derived by inference from other beliefs directly based on perception. Counterfactuals 1 and 2*, for example, would then be known independently of experience, but they would fail the second condition. Propositions 4, 5 and 6, on the other hand, would be known independently of experience and they would also plausibly meet the second condition, so knowledge of 4, 5 and 6 would be counted as a priori. There are however a number of problems remaining. A first, maybe minor, problem with this proposal is that it gives bizarre results if we admit that some necessities do not iterate. Suppose there is a proposition *p* such that necessarily *p* is true, but it is not the cases that necessarily, necessarily *p*. In such case, on the present view you could know a priori that *p*, but you could not know a priori that necessarily *p*, even if you had a proof of the latter claim which was completely independent of experience.

A further problem is that MAP might classify some of what we take to be, after Kripke, a posteriori knowledge of necessary truths as really a priori. Consider the belief that water is H_2O, or the belief that Hesperus equals Phosphorous. These beliefs need not be based on experience in the narrow sense, and their content is either necessarily true or necessarily false, so they will end up being a priori justified for someone. This wouldn't make a posteriori knowledge of necessities impossible, since it could be possible for someone to have evidence directly dependent on experience, in the narrow sense. I am not sure how bad this consequence is. Presumably the cost might be deemed low enough to be paid in exchange for saving the role of the a priori/a posteriori distinction.

The third problem is that MAP, as we noted before, rules out the possibility of the contingent a priori. Of course all cases of putative contingent a priori knowledge are controversial; but, on this view, we do not need to look at the cases, for we know from the start (a priori, perhaps) that no such case is possible. Such a stance is question-begging, unless an explanation is provided of the link between the modal status of the proposition and the epistemological status of our belief in it. This leads us to what I see as the most important problem with MAP.

The fourth and last problem is that it is not clear why we should take the notion of a priori that MAP delivers to have any epistemological interest. If we can know contingent propositions independently of experience, in the same sense in which MAP talks of

independence from experience, this seems a good reason to classify that knowledge together with the knowledge of necessary truths which is similarly independent of experience. Compare this case with a case of a posteriori knowledge of necessities, in which I come to know that water is H_2O, e.g., on the basis of inference from my own perceptually justified beliefs, performing all the relevant experiments. My knowledge that water is H_2O meets the second condition in MAP, but not the first. However, we are not tempted to classify this piece of knowledge as a priori; the modal status of the proposition seems epistemologically irrelevant. My knowledge of 1 (if two marks had been nine inches apart, they would have been further apart than the front and back legs of an ant), meets the first condition, we are now assuming, but not the latter. Why shouldn't we say that this is a case of a priori knowledge of a contingent truth? Why is the modal status of the proposition epistemologically relevant in this case? We are left without an explanation, as far as I can see. Plantinga [1993] writes: "the question here is whether the term 'a priori knowledge' expresses the concept of knowledge independent (in the right way) of experience, or whether it expresses a stronger concept: the concept of knowledge independent of experience accompanied by the conviction that what is known is necessary. This is the sort of question to which there may be no answer; the thing to do is to note both concepts but bracket the question which concept is expressed by the term" (p.107). We have seen above that using the subject's conviction that the proposition is necessary gives unwanted results. But perhaps we may apply Plantinga's irenic attitude to the question whether the term 'a priori' expresses a concept that requires, to be properly applied, only that the first condition in MAP is satisfied, or a concept that requires also that the second is satisfied. Insofar as the usage of the theoretical term 'a priori' is sufficiently confused, I may agree with Plantinga that there is no determinate answer to the question. But this still leaves room for the further question: which of the two concepts is more fruitful to employ (if any)? Clearly the conjunctive concept is at disadvantage here, because it seems to create an unnatural category[60]. It would be just like considering from the point of view if biology to propose to use the term "mammal" to express the concept of being a mammal and not living underwater. We can certainly stipulate that use of the term, but what is its purpose? It would be equally illuminating, from the

[60] This can be seen also by considering a case discussed by Williamson in a similar context (Williamson [2007] p. 51, fn.3). Suppose someone knows in a fully empirical way that water is H_2O, and knows independently of experience that the meter bar in Paris is one meter long. Neither of the two beliefs would be a priori on this definition, but their disjunction would be; it would be independent of experience because of one disjunct and necessary because of the other.

epistemological point of view, to create a category of knowledge which satisfies the following definition:

Feline A Priori(FAP): A subject has a priori justification for a proposition *p* iff i) the subject has a justification for *p* independent of experience and ii) *p* is a proposition about cats

In absence of an explanation of the epistemological relevance of ii in FAP, I think we should say that FAP does not provide an epistemological distinction at all, and *a fortiori* not an interesting one; and we should say the same about MAP. It might be thought that there is some further explanation of the difference in epistemological significance between a proposition having a certain modal profile and its being about, say, cats. I will look at one possible way of spelling out that difference in the next section.

3.4. Understanding and Justification

The traditional way of solving the problem of demarcating experiences epistemically relevant to a justification from those that are only causally relevant, as I mentioned earlier in connection with Hawthorne's example, is to appeal to the evidential/enabling distinction, and to spell out this distinction in turn in terms of the role of understanding in generating a justification. In the case of the a priori, understanding is sufficient for justification (in a sense to be specified), and therefore we may regard the role of experience in providing understanding as merely enabling.

Let us give this idea a somewhat more precise form:

Understanding-Justification A Priori (UJP): A subject has a priori justification for a proposition *p* if and only if the subject has a justification for *p* which is based on the subject's understanding of *p* and reasoning independent of experience.

Read 'reasoning' in UJP in such a way to include, perhaps as a limiting case, the process of considering a proposition and endorsing it (one may also think of that as inference from an empty set of premises). Let us note that endorsing UJP does not amount to endorsing the existence of epistemic analyticity, in any of sense discussed in chapter 2. The idea of epistemic analyticity is that understanding *by itself* is sufficient for either belief or justification. To see the difference, it might help to compare UJP to the following two somewhat related principles, discussed in Williamson [2007]:

> KUt' Whoever knows *every vixen is a vixen* in the normal way does so simply on the basis of their grasp of the thought
>
> (…)
>
> KUl' Whoever knows 'Every vixen is a vixen' in the normal way does so simply on the basis of their understanding of the sentence (Williamson [2007], pp. 130-131)

Williamson rejects the two principles precisely on the ground that the normal way to come to know the proposition involves some form of logical or semantic competence that goes beyond what is required to understand the sentence, or entertain the thought. I take such considerations to be conclusive against KUt' and KUl'. But the rejection of these principles leaves open the possibility that whoever knows 'Every vixen is a vixen' in the normal way[61] does so a priori, in the sense of UJP.

The main point of criticism that I will raise against UJP is that problems similar to the ones we met so far will reappear in the understanding of what it is for some reasoning to be independent of experience. That the qualification "independent of experience" should be attached to "reasoning" in UJP in the first place can be seen as follows. Firstly, note that if the qualification were not there, all the cases we considered so far would come out as a priori. In none of the cases perceptual experiences, as we have seen, play any direct role in the justification of the beliefs. They play at most a role in shaping the reasoning processes involved. Secondly, and more generally, a subject could acquire a reasoning pattern, such as moving from claims of the form 'x is F' to claims of the form 'x is G', through being exposed to experiences of the right sort. In the sort of case imagined, one could acquire the reasoning pattern through the same experiences that would inductively justify the claim "all Fs are Gs" – although the subject need not form that belief, or even possess the concept ALL. The subject could then form, about some particular object, a belief of the form 'if that is F, then it is G'. Clearly, that belief could not constitute a priori knowledge, or a priori justified belief. So we need to restrict UJP to only allow reasoning independent of experience, and therefore we need a way to specify the independence from experience requirement for reasoning.

One could think that appeal to the modal status of the propositions is relevant here, and that the modal status of an inference provides a way to clarify the sense in which the

[61] In the light of previous considerations, it might be that many mathematical truths are not known in the 'normal way', since the subject, at least the relatively mathematically unsophisticated subject like me, will rely heavily on testimony. But it might still be the case that simple mathematical and logical truths, and the content of judgements about most philosophical thought experiments, are known a priori. This would be a good result, I take it, for the defender of the usefulness of the distinction.

reasoning involved can be independent of experience. It could be said that reasoning bringing to a necessarily true conclusion delivers a reliably true result, *because* the proposition involved is necessarily true; which would not be the case in the other problematic cases considered above. This way, we would also have a tie between necessity and a priority. For when contingent truths are involved, the disposition to move from certain premises to a certain conclusion depends for its reliability on the way it was acquired; in the necessary truths cases, it does not, since it is determined by its content that the resulting belief will be true.

However, further reflection shows that the modal status of the proposition is irrelevant, and the argument to the contrary I just went through is flawed. The fact a proposition is a necessary truth does not automatically guarantee that any belief in that proposition was formed in a reliable way. This is indeed a problem for certain formulations of an externalist constraint, such as the following (intending to capture a necessary condition for knowledge):

Safety: A belief that p is safe iff in all (or most) close possible worlds in which the subject has the belief that p, p is true. [62]

I cannot do better here than quoting Sainsbury's (Sainsbury [1995]) very clear explanation of the problem:

> Suppose I guess correctly that 193 + 245 = 438. I do not know that 193 + 245 = 438, because my belief does not issue from a reliable mechanism. This cannot be brought out by pointing to similar possible worlds at which 193 + 245 sum to something other than 438, since there are no such worlds. Arguably, it can be brought out by pointing to possible worlds in which I exercise essentially the same capacities and yet arrive at a falsehood. A lucky guess is not a proposition which might easily not have been true, but a way of reaching a belief which might easily not have delivered a true one. (Sainsbury [1995] p. 595)

In other words, if I believe a necessary truth, my belief is such that it could not be wrong, and one might think that this is an epistemically important feature; but the example shows that it is not. It is compatible with that situation that my belief is not only irrational, but also formed in an unreliable way (in a pre-theoretic sense). Similarly, going through a valid inference can be an irrational and unreliable reasoning process. Sometimes the epistemological distinction between contingent and necessary truths is spelled out as follows: in order to know that a contingent proposition is the case, you surely have to look at the way the actual world is; after all, the proposition could be true, and it could it be false. You need to look at the actual world

[62] Compare Pritchard [2005] p. 71

to see which one it is. Talk of 'looking and seeing' is of course metaphorical; a more neutral way of putting the point would be saying that you need interaction with the actual world in some form; and all interaction with the actual world, it might be thought, either *is* perception of goes through perception (like testimony). But if a proposition is necessarily true (or false), you do not need to look; in some cases at least, you come to know it (or its negation) independently of the interaction with the actual world. After all, the actual world could not falsify your belief. However, we must be careful not to neglect the fact that although the world cannot, as it were, outstrip the limits of necessity, our beliefs can. We can believe necessary falsehoods (and in that case, looking at the actual world might surely help). So we need a reliable way of forming beliefs even in this kind of case, in order to know something. But, as a matter of fact, we acquire our belief-forming method through interaction with the actual world, and this is epistemically crucial; or so I will argue.

Any interesting account of reliability should take care of the problem of spelling out non-trivial conditions for a belief in a necessary truth to be reliable[63]. Sainsbury's way of solving it[64] is through this modified definition of safety

Mechanism-Safety: A belief is safe iff it is produced by a mechanism which could not have easily produced a false belief.

This will serve as a reasonable approximation for the sort of requirement our judgements must meet, in order to be reliable, in the case of necessary propositions. But the account makes essential use of the notion of the "mechanism" by which a belief is produced. Therefore, we are presented again with the problem of individuating the relevant mechanism (or method; or process; I will treat all these as synonyms for present purposes) by which we come to believe paradigmatic philosophical truths. Williamson argues that we should individuate it in a way that makes reference to how the disposition to judge is acquired, and in particular to the experiences that we undergo in the process of acquiring it. The relevant process, everyone can agree, is a sort of reasoning process that does not take direct perceptual inputs, and this

[63] Careful defenders of reliabilism have been largely aware of the issue for a long time (see e.g. the discussion in Goldman [1986] ch. 3). I am here looking at Sainsbury's account just as an example. Weatherson [2004] develops a proposal along similar lines, which has the advantage of not involving the notion of a method, but requires to deny that beliefs are individuated essentially by their content. Nothing in the examples developed later turns on this choice, as far as I can see.

[64] See Sainsbury [1997]

process must be the expression of a reasoning capacity[65]. In case this capacity is epistemically in good standing (e.g., it allows the reasoner to move from justified premises to justified conclusions), I will call it a reasoning *ability*. However, what constitutes the possession of a reasoning ability is partly its origin. Moreover, typically the origin of a human reasoning ability is tied to the experiences of the subject. Williamson illustrates this point through the following example

7) If two marks had been nine inches apart, they would have been more than nineteen centimetres apart.

We are asked to consider a situation in which the subject judges 7 to be true, correctly, by visually imagining the two lengths in inches and in centimetres, and comparing them in imagination. The subject does not need to use memory of any specific episode of judging distances. But then, Williamson argues:

> Whether my belief in [7] constitutes knowledge is highly sensitive to the accuracy or otherwise of the empirical information about lengths (in each unit) on which I relied when calibrating my judgments of length (in each unit). I know [7] only if my off-line application of the concepts of an inch and a centimeter was sufficiently skilful. Whether I am justified in believing [7] likewise depends on how skilful I am in making such judgements. My possession of the appropriate skills depends constitutively, not just causally, on past experience for the calibration of my judgments of length in those units. If the calibration is correct by a lucky accident, despite massive errors in the relevant past beliefs about length, I lack the required skill. (Williamson [2007] pp. 166-167)

I will argue that Williamson's assessment of this case is correct, and it extends to paradigmatic cases of philosophical knowledge, such as 4-6, the truths about knowledge and belief considered in section 3.2. My argument requires considering the following thought experiment:

Philosophy Oscar

[65] Not any change in intentional mental states will constitute reasoning. If I am thinking about the weather, and then I hear a strong noise, my thoughts will change, but not as the effect of reasoning. However, it turns out to be extremely hard to say what relation exactly has to hold between two thoughts in order for the process of moving from one to the other to count as reasoning. Wedgewood [2006] proposes that such a process must be causally influenced by the normative relations between the thoughts. If something along these lines is correct, then perhaps one need not distinguish, as I am doing, between a capacity for reasoning and a reasoning ability. I would then reformulate the arguments that follow addressing directly the capacity of reasoning.

Suppose Oscar is a fairly normal subject, who has studied some philosophy, but has never reflected about the relation between knowledge and belief in particular. He is asked what he thinks of the claim that if someone knows that something is the case, then she has to believe that it is the case. After a short reflection, he forms the belief that the claim is obviously true. Suppose also Oscar has a twin, call him TwinOscar (whether the twin is actual or merely possible does not matter). TwinOscar has also studied some philosophy. Unfortunately, however, he has read a lot about the relation between knowledge and belief, and he has come to hold the view that knowledge does not entail belief (I will assume such a view to be false; of course the example could be varied if needed). He then found apparent confirmation of his view in the actual world; he often judged people he met to know some things they did not believe, for lack of adequate confidence. But at some point TwinOscar came to realize that such a view was very unpopular among philosophers, and he formed a desire that he did not hold the view, for that, say, would help him in his career. Luckily, a friend of TwinOscar is a neurologist, and she happened to develop a pill such that it produces in whoever assumes it the belief that knowledge does entail belief, and it cancels at the same time any memory of taking the pill, and any memory of having doubted in the past that knowledge entails belief. So TwinOscar takes the pill; a few seconds later, He is asked what he thinks of the claim that if someone knows that something is the case, then she has to believe that it is the case. After a short reflection, he forms the belief that the claim is obviously true.

Now, recall our examples 4-6:

4) It is necessary that whoever knows something believes it.
5) If Mary knew that it was raining, she would believe that it was raining.
6) Whoever knew something believed it.

TwinOscar understands each of 4-6, and, at the end of our story, he could clearly judge them to be true. The way he forms the belief that 4-6 are true can be subjectively indistinguishable from the way Oscar forms the belief. However, while Oscar knows 4-6, TwinOscar fails at that point to know any of 4-6 (he might of course come to know them later, when his judgements are calibrated again through his experience of interacting with other human beings). The best explanation of the difference seems to be that Oscar's capacity to apply the relevant concepts, unlike TwinOscar's, was appropriately formed through his past

experiences, and therefore it constitutes an ability. The only alternative explanation I can think of is that TwinOscar's thought process in this case does not constitute reasoning at all[66]. I find this implausible, and I am doubtful that the case cannot be developed in such a way to make this too implausible to be even a candidate explanation of what is going on. But I do not need to convince the reader of that. Suppose Oscar is capable of reasoning while TwinOscar, in that specific situation, is not. Then again, we need an explanation of this fact, since Oscar and TwinOscar are physically and phenomenologically indistinguishable. Whether a subject counts as reasoning, the only explanation would be, depends on the relation between the subject and her past experiences.

A conclusion we have reached at this point is that possession of the sort of ability in applying a concept required for philosophical knowledge depends constitutively, at least in some cases, on the way the ability was formed, as Williamson claimed. The more general lesson of the case is that the notion of a reasoning ability cannot be understood purely internalistically; the ability depends, for its normative status, on factors that are beyond the (present) internal state of the subject. We will come back to this consideration, and to further questions about the notion of an ability, in Williamson's sense, in the next chapter.

There are an objection and a question which I need to address at this point. The objection has to do with the possibility of defining an internalist notion of a priori justification which escapes the sort of argument just given. The question is how smoothly the foregoing considerations apply to other paradigmatic cases of a priori such as mathematics and logic. Let me address the objection and the question in turn. The objection would start by contesting the point made by Williamson in the case of 7, that whether he is "justified in believing [7] likewise depends on how skilful [he is] in making such judgements." Since we saw that subjectively indistinguishable subjects can have different skills, here Williamson is really assuming an externalist notion of justification. But this is not, the objection would go, the sense of justification we were interested in, for we were after justification in a more internalist sense. A belief is justified in this sense when it is rational, roughly, from the subject's point of view, and it has not been shown that there is no interesting notion of justification independent of experience in this sense.

The issue of whether such a notion of justification is in itself viable is beyond the scope of the present work. So, although this is controversial, I will assume for the sake of the argument that the notion is not defective in itself, so that the objection can get off the ground. The problem with the objection is that the notion of reliability was brought in, initially, to

[66] Wedgewood's account of reasoning mentioned in fn. 65 would perhaps have this result.

clarify the relevant sense of independence from experience in the first place. The problem was precisely that some paradigmatic cases of a priori justification seemed to be exactly similar, from the subject's point of view, to cases we did not wish to classify as a priori. The suggestion was then that in the a priori cases understanding and reasoning gave the subject a superior sort of reliability. Then I argued against that suggestion. On the line of objection that I am considering, it was never a good idea to understand the difference in terms of reliability. But then I have no idea of how the distinction could be characterized. Section 3.2 above argued against some initially plausible proposals. It is true that I have no general argument to the effect that there is no interesting notion of a priori justification in the internalist sense of 'justification'; but I see no reason to think that there is one.

The question as to whether the considerations we developed for the case of 4-6 apply to other paradigmatic, and even more so, cases of a priori knowledge is a very good one. Williamson tentatively suggest that they do; he writes:

> In a similar way, past experience of spatial and temporal properties may play a role in skilful mathematical 'intuition' that is not directly evidential but far exceeds what is needed to acquire the relevant mathematical concepts. The role may be more than heuristic, concerning the context of justification as well as the context of discovery. Even the combinatorial skills required for competent assessment of standard set-theoretic axioms may involve off-line applications of perceptual and motor skills, whose capacity to generate knowledge constitutively depends on their honing through past experience that plays no evidential role in the assessment of the axioms. (Williamson [2007] pp. 168-9)

I admit that nothing more than Williamson's cautiously tentative conclusion has been, or will be, defended here. The considerations above seem to indicate at least a prima facie difficulty in spelling out a sense of a priori that could cover mathematical and logical reasoning without covering too much. But the reason I am interested in the issue of the a priori/a posteriori distinction is, just like it was for Williamson (at least in Williamson [2007]), my interest in philosophical methodology. So I would be very happy to have established that no viable way of spelling out the distinction leaves paradigmatic cases of philosophical knowledge on the a priori side without also classifying as a priori large chunks of what would have been thought to be straightforwardly empirical knowledge of the world. Since 4-6 are paradigmatic pieces of philosophical knowledge, assuming the arguments above are successful, that aim has been reached. Moreover, it seems easy to see how parallel considerations would extend to many other cases of philosophical thought experiments.

Moreover, I wish to indicate a line of research in philosophy of logic and mathematics that is consonant to the present approach. Such a line of research is best exemplified by the work of Penelope Maddy (e.g. Maddy [2007]). The general thrust of Maddy's view is that abstract concepts such as the concept of set, or the concept of object, reflect real features of the structure of the world. While such concepts, as recent research in cognitive psychology suggest, might be largely innate (see e.g. Maddy [2007] pp. 245-270), we learn how to apply those concepts primarily through our perceptual interaction with the world. There are complex exegetical and substantial issues looming large here, but it should be clear how such an approach can naturally be combined with the present perspective on the a priori.

Conclusion

In this chapter, I have discussed three kinds of replies to the challenge posed by arguments given by Williamson [2007] and Hawthorne [2007] against the distinction between a priori and a posteriori. Williamson and Hawthorne argue that, on the background of the rejection of epistemic analyticity motivated by semantic externalism, there is no way to spell out the distinction which provides an extension approximately similar to what philosophers thought the extension to be. The first reply I discussed, defended by Jenkins [2008b], is that the arguments can be defused by using the notion of grounding. But it turned out that there is no understanding of that notion such that the notion does the required work. A second kind of reply consists in making some restrictions on the modal profile of the proposition which are known a priori; this turned out to be both problematic in a number of cases and unmotivated in general. Finally, I considered the proposal that a priori justification is justification based purely on understanding and some reasoning. But I argued that this reply needs to restrict the reasoning involved to reasoning independent of experience, and this will exclude at least some cases of beliefs which were traditionally thought to be paradigmatically a priori, in particular philosophical beliefs.

This chapter concludes my case in favor of anti-exceptionalism. In the next chapter I will try to develop a positive proposal as to the origin of some philosophical knowledge, by also developing some considerations already presented in the last section of this chapter.

Chapter 4
Thought Experiments and the Application of Concepts

Introduction

In this chapter, I give a positive account of one aspect of philosophical methodology, the use (or, a certain use) of thought experiments. I do not want to exaggerate the importance of this aspect; there are many other things philosophers do besides considering thought experiments, and when they do consider thought experiments, they might do so in different ways and for different ends (I will come back to this at the beginning of the next chapter, where some confusion between different uses of thought experiments might be relevant). Still, philosophers do use thought experiments in the way I am going to describe; and many think there is something problematic about that, or at least something that requires explanation. I am going to try to describe such a use of thought experiments, to articulate carefully what is thought to be problematic about it, and to provide an explanation of what goes on which shows that there is no problem after all.

What is a thought experiment, exactly? Robert Brown, in the related entry in the Stanford Encyclopedia of Philosophy, writes that "Thought experiments are devices of the imagination used to investigate the nature of things."[67] Roy Sorensen defines a thought experiment as "an experiment (…) that purports to achieve its aim without the benefit of execution."[68] He refers back to the following definition of experiment:

> An experiment is a procedure for answering or raising a question about the relationship between variables by varying one (or more) of them and tracking any response by the other or others. For the sake of simplicity, most experiments are designed around two variables. The one you directly manipulate is called the independent variable and the one you try to affect indirectly through these manipulations is the dependent variable. (Sorensen [1992] p.186)

[67] Brown [2007]

[68] Sorensen [1992] p.205

These definitions are of course vague, and I will argue that they can be read in an extremely liberal way, so that any case of hypothetical reasoning will count as a thought experiment. I do not think we can, or should, provide more precise or more restrictive definitions. Of course it is quite common to point at examples of the phenomenon for better elucidation. Typical examples of thought experiments in philosophy, which are our main concern here, include Gettier cases, Putnam's twin earth case, Jackson's Mary, trolley cases, and so on. What is common among these seems to be that we are asked to imagine a certain situation, often an unusual or even far-fetched one, and then we are asked to judge about a certain feature of the situation; does the subject know? Does the subject's word 'water' refer to H_2O? Does Mary learn something new? Should you throw the switch to divert the trolley? And so on. The answer to this question is then used as a premise to reach some interesting philosophical conclusion. One risk of this strategy is that it could lead us to overlook certain similarities between the thought experiments used in philosophy and more common acts of thought. Paradigm cases are sometimes misleading, as we know. For instance, suppose we first identified planets as some celestial bodies (certain 'stars', as we might have said) which were moving in an apparently irregular and somewhat surprising way in the night sky. In that situation, the suggestion that the ground on which we were standing was itself part of a planet might have seemed implausible; it did not fit the paradigmatic cases. We could be making a similar mistake in thinking that thought experiments in philosophy are radically different from more ordinary specimens of hypothetical reasoning; we might be failing to look at the epistemic ground on which we stand. A related point is that we should not focus exclusively on the most hard and controversial cases of philosophical thought experiments, which of course tend to attract most of the attention.

Of course examples of everyday hypothetical reasoning abound. I might imagine that I dropped my cup of coffee, and judge that in that case the cup would have fallen to the ground and spilled its content on the floor. Or I could imagine that there is a law in the UK which prohibits comparing philosophical thought experiments to any kind of reasoning occurring outside philosophy, and judge that in that case I would be breaching the law right now. Are these thought experiments? Williamson [2007] suggests that there is no essential difference between thought experiments in philosophy and ordinary counterfactual judgements, e.g. 'if I dropped my cup of coffee, it would fall on the ground and spill its content'. The latter sentence fits, although somewhat loosely, Brown's definition. I might be using a device of the imagination to investigate the nature of my cup of coffee, or perhaps the nature of physical

laws in our universe. Similarly, the reasoning described fits Sorensen's definition; I am affecting (in imagination) the independent variable of my holding the cup, and observing an effect on the dependent variables constituted by the cup and its content.

Sorensen [1992] and Williamson [2007] defended the claim that there is nothing exceptional about the use of thought experiments in philosophy. The forms of reasoning, and the sort of cognitive capacities, involved, are the same as those we routinely use outside philosophy, and the epistemic evaluation appropriate for the resulting judgements is also the same. Of course this is in itself not a very strong claim, because it is far from clear what forms of reasoning we do and do not use outside philosophy. I will try to give more substance to the claim later, as well as defending it. I need to make explicit two assumptions which I am making at this point, which were defended in chapters 1 and 2: first, judgements about thought experiments in philosophy do not rely in a crucial way on 'intuitions' or 'seemings'; at any rate, not more than any other judgement. Secondly, possessing the relevant concepts is not sufficient to produce the judgement or to have justification for the judgement.

I take the main challenge I am addressing in this chapter to be providing an answer to the following questions: given the foregoing assumptions, can we gain knowledge from thought experiments? And, if so, *how*? The answer I wish to give is along the following lines: we gain knowledge through thought experiments by using a capacity for making judgements about possible situations, and that capacity just is a particular application of our general capacity to apply concepts.

However, before getting to this main task, I have to address an issue about thought experiments which has been recently the object of some discussion, and which has considerable independent interest. The issue is what description we should give of the logical form (in a sense to be explained) of the reasoning involved in thought experiments. I will defend a particular view on the issue, which will help frame the following discussion. In the second section I will articulate a worry about the use of thought experiments which is the main focus of this chapter. In the third section I will provide an answer to the worry, and in the fourth I will explore some consequences of that answer.

4.1. Thought-Experiments and "Logical Forms"

One way to put the question I am going to discuss in this section is: what is the best way of representing the logical form of thought experiments? But that question is not as clear as one might wish. It is not clear what 'logical form' refers to in this context, and it is not clear from what point of view we are to evaluate different representations of it. Are we trying

to describe the way subjects actually think, at some level, when they make use of thought experiments? Or are we trying to describe the way subjects should think, at some level, when they entertain thought experiments? I take the task to be mainly the latter. Of course considerations related to human psychology will still not be irrelevant, assuming ought implies can. Therefore I take the aim to be that of showing how an idealized (human) subject could employ thought-experiments in her reasoning processes. A further way in which the question could be clearer is by indicating the scope of the cognitive activity we are considering. On one extreme, we might be asking about the mere act of entertaining a thought experiment; presumably that requires no more than understanding the text, in the relevant interpretation. On the other extreme, we might be asking about any kind of reasoning that involves thought experiments. I will take a middle ground here. The former alternative would perhaps be attractive to someone who thinks that thought experiments are used to raise questions, more than to provide answers, or that they are purely rhetorical means of persuasion. Thought experiments are often used in these ways, but that is not the sort of use which is discussed here. What we are discussing here is primarily the following kind of use of a thought experiment: you consider a hypothetical scenario, you draw a judgement about it, and then you draw some philosophical conclusion from the judgment; that is, from the content of the judgement (perhaps in combination with other premises). Most of the discussion has focused on a specific use of judgements about hypothetical scenarios, that of refuting a necessity claim. However, there are other possibilities. A judgement about a case can be use to support a philosophical theory, e.g. by inference to the best explanation, if the theory predicts the correctness of the judgment. Consider for example the following argument. In a twin earth scenario, XYZ is not water, but H_2O is. Since XYZ in that scenario has all the superficial features of water, and plays the same social role as water actually does, the best explanation of the fact that H_2O, and not XYZ, is water in the twin earth case, is that water is *necessarily* H_2O. On the other hand, we certainly cannot try to formalize all possible uses of thought experiments in reasoning. The assumption is that, in at least many typical uses, there is some interesting common core, and we are trying to capture that common core.

Since much of the discussion has focused on one example of a thought experiment, the Gettier cases, it will be useful to focus on those here as well. There is no assumption that all different thought experiments, or even just philosophical thought experiments, will fit the proposed accounts without modifications. However, an assumption I am making (together with almost everyone else in this debate) is that, in at least a significant amount of cases, there will be enough similarity for the analysis of Gettier cases to provide some insight on the

logical form of the reasoning involved in using thought experiments. The similarity might not extend beyond the aspect of logical form. A Gettier case is here defined as a short text which is intended by its author[69] to provide a possible case of justified true belief that falls short of knowledge. Here is an example:

> Tom knows that Federer is playing the final of Wimbledon. He turns on the TV and sees Federer hitting an ace on the match point of a game in Wimbledon's central court. Tom therefore forms the belief that Federer has won Wimbledon this year. In fact, Federer has won, but because of a technical problem the game shown on TV was not this year's final, but last year's final[70].

Informally, it seems quite clear what is going on when we consider such a case as a counterexample to the theory that knowledge just is justified true belief. We are supposed to reason along the following lines: Tom's belief is justified; he has no reason to suspect that the game shown is not live, and everything he has seen is consistent with his information. His belief is also true. Yet, he does not know. Therefore, knowledge is not justified true belief.

Nevertheless, there is controversy on what is the best way to represent precisely what is going on here. In particular, I have in mind the debate between Williamson [2007] and Ichikawa and Jarvis [2009]. All parties to this debate agree on a number of things; crucially, they agree that the reasoning using the thought experiment can be represented in argument form. This is a rather substantial assumption; however, it should be clear that firstly, as I said before, there is no assumption that all thought experiments, and their use, will conform to this model. Secondly, there is no assumption that the *only* thing which is interesting about thought experiments is their use in arguments. These points should make the assumption look less demanding. To represent the use of a Gettier case in argument form, let us use the following abbreviations

GCx,p = x stands in the relation described by the story to a proposition p

JTBx,p = x has a justified true belief that p

---

[69] In a different sense, a Gettier case is a successful attempt at providing a possible case of justified true belief without knowledge. I prefer this formulation because authors who think that all such attempts fail, since knowledge is justified true belief, could still want to talk about Gettier cases as an interesting class of situations (see e.g. Weatherson [2003]).

[70] This is just an updated (from the tennis point of view) version of a case described in Dancy [1985], p. 25

Kx,p = x knows that p.

We can now represent the claim that this thought experiment is supposed to disprove as follows:

C Necessarily, for all x and all p, $JTBx,p \leftrightarrow Kx,p$

Moreover, it is agreed that one of the premises will say that the case described is possible. We can represent it as

1) Possibly, there is an x and there is a p such that (GCx,p)

Now, Williamson [2007] argues that the second premise cannot be

2) Necessarily, for all x and p, $(GCx,p) \rightarrow JTBx,p \wedge \sim Kx,p$

For, although 1) and 2) clearly entail the negation of C, he thinks that 2 is false. For example, suppose that someone satisfies the description given in the above text, but he has a friend who was watching the match live in Wimbledon, and that friend has called him and also told him that Federer has won. Then his belief is knowledge after all. Or imagine that he has read in a reliable newspaper that Federer would play Nadal in the final, and he is instead seeing Federer playing Roddick; he can distinguish the two, but he irrationally distrusted the newspaper anyway. Then his belief is not justified after all. Williamson claims that the ways things could go wrong, for one who is arguing against C, are so numerous that it would be impossible to present a variant of the story that prevents them all. But it would also be unnecessary, because we can use as a second premise the following counterfactual:

2*) For some x and p, $(GCx,p) \boxed{}\!\!\rightarrow$ For all x and all p, $(GCx,p) \rightarrow (JTBx,p \wedge \sim Kx,p)$

This is Williamson's preferred formalization of the English counterfactual: "If a thinker were Gettier-related to a proposition, he/she would have justified true belief in it without knowledge"[71]. The argument form 1 and 2* to the negation of C is valid, and it avoids the problem mentioned for 2.

Ichikawa and Jarvis defend the aptness of (something close to) the original premise 2. They claim that its apparent inadequacy is due to an incorrect way of thinking about the story presented in the Gettier case. We should not identify the story with the text; rather, the story is what we get when the text has been interpreted, in the way we standardly interpret works of fiction. They also argue against Williamson's alternative proposal. I will discuss two objections that they raise against Williamson, and I will argue that they both fail. I will then

[71] Williamson [2007] p. 195

briefly go back to consider the proposal advanced by Ichikawa and Jarvis. The first objection that Ichikawa and Jarvis raise is that Williamson's way of representing the argument makes it a posteriori, for evaluating a counterfactual such as 2* requires one to employ various pieces of world-knowledge which are clearly empirically justified, if justified at all; as a result, the crucial premise would not come out a priori, and they regard this as a disadvantage. However, I have argued in chapter 3 that Williamson's objections to the epistemological significance of the a priori/a posteriori distinction are independently well motivated. So I will not spend time here on this objection. The second objection they raise is more interesting in the present context. The objection is that premises of the form of 2* will often be false (exactly how often, is not completely clear, and we will come back to this). They will be false when the nearest (in terms of possible worlds) realization of the Gettier text, which is the antecedent of the counterfactual, is one which falsifies the consequent; in particular, this will happen whenever the Gettier text is *actually* true of someone, but it is not actually true that the person lacks knowledge, or it is not actually true that she has justified true belief. In Ichikawa's [2009] example

> Suppose that one's thought experiment is given thus:
>
> *At 8:28, somebody looked at a clock to see what time it was. The clock was broken; it had stopped exactly twenty-four hours previously. The subject believed, on the basis of the clock's reading, that it was 8:28.*
>
> This should be recognizable as a prototypical Gettier description.
>
> Now consider a world in which that description is true, but where the subject knew in advance that the clock had stopped exactly twenty-four hours previously. In that world, the Gettier text is true but misleading: its subject knows. So [the relevant counterfactual] is false in that world [footnote attached: this is so on a standard Lewis-Stalnaker semantics for counterfactual conditionals, and on any account on which A $\wedge\sim$ C entails the falsity of A $\Box\!\!\rightarrow$ C]. Someone running the Gettier argument in that world, then, relies on a falsehood, even if he is innocently ignorant of the person who happens to render his counterfactual false. Relatedly, in running the Gettier argument, one commits oneself to being in a world *not* positioned in a way that falsifies [the relevant counterfactual]. I take these implications to be implausible. (Does one fail to know the Gettier conclusion by virtue of there being someone in his world who satisfies the text in the wrong way?) (Ichikawa [2009] p. 437)

As Ichikawa [2009] notes, Williamson [2007] already contains some discussion of this kind of objection[72]. Williamson has two lines of reply. He says that the quantifier in the

[72] As explained in Ichikawa [2009], Ichikawa and Jarvis [2009] was targeted at early presentations of Williamson's view, in particular Williamson [2005].

antecedent of 2* might be restricted by the context so that it excludes some cases from the domain of quantification, even though they are actual or they occur in nearby worlds. Still, Williamson admits that in some cases the counterfactual will turn out to be false. He argues that this is not a problem for his view. First, he claims that when someone is shown that the text of a thought experiment is actually realized in such a way that it makes her intuitive judgment false, the correct reaction is simply to modify the example so as to take care of the problem. Williamson concedes that we might often be tempted to insist that this is not necessary, but he puts that down to a general tendency to fail to admit mistakes. That we have this tendency is certainly true, but something more needs to be said. Williamson also gives an example that suggests an additional reply:

> ..suppose that someone says 'Every man in the room is wearing a tie'; I look around, see a man not wearing a tie, misidentify him as Dave (who is in fact wearing a tie), and say 'Dave isn't'. When it is pointed out to me that Dave is wearing a tie, I deceive myself if I insist that my original reply was correct because the man whom I had in mind was not wearing a tie; that was just not the 'counterexample' which I actually presented. I spoke falsely when I said 'Dave isn't'. (Williamson [2007] p.201)

The example not only suggests a way in which one might be tempted to resist the need to modify one's claims; it also illustrates, I think, a certain sort of value that there can be in a false claim. The subject's claim in the example can be informative, drawing the attention of the party to the relevant counterexample to the generalization under discussion, even though it is false. As a result of the false claim that Dave is not wearing a tie, people may realize that someone else, the man I point to saying 'Dave', is not wearing a tie. We should say something similar, I will argue, for Gettier cases which are realized in an unintended way. I will argue that there is plenty of room to explain why the judgements about those cases still have a lot of epistemic value, and why we are reluctant to think they are false. To illustrate this point, it will be useful to consider a couple of thought experiments about thought experiments:

Case 1. Suppose a subject, call him John, has his first encounter with a Gettier case reading Williamson's description of a real life Gettier case. Williamson claims that he gave his students the false information that the only power-point presentation he has given in his life was a failure, while in fact he never gave a power-point presentation, successful or otherwise. He then made sure that the students could clearly see that if his only power-point presentation was a failure, then he has never given a successful power-point presentation. On the basis of Williamson's testimony John judges that the "victims" of this machination did not know that Williamson has never given a successful power point presentation, although they

had a justified true belief to that effect. However, suppose also that Williamson's description of the actual Gettier case is mistaken in the following way: the day before the lecture, Williamson had told of his intention to create a real life Gettier-case to a colleague; unbeknownst to Williamson, the colleague told of this plan to some of the students, and the whole plan, with its details, became common knowledge (moreover, we can imagine that it is common knowledge among Williamson's students that he has never given a power-point presentation at all). So after all the students knew that Williamson has never given a successful power point presentation.

Case 2. Suppose a subject, call her Jane, is introduced to Gettier cases by the following example: Smith is told by an apparently reliable and honest mathematician that Fermat's last theorem is false. The mathematician actually wishes to deceive Smith. But at the same time, another mathematician, unbeknownst to both Smith and his informant, has just proven that Fermat's last theorem is false. Jane also mistakenly believes, based on an apparently reliable testimony, that Fermat's last theorem is false. She judges that Smith, in the story, has a justified true belief which is not knowledge.

There are four points, which I want to make about these two cases. I take all four points to be entirely uncontroversial, but it is important to keep them in mind.

1) Everyone, or at least anyone who thinks thought experiments can figure in an argument, should admit that John and Jane form some false beliefs and they make use of them in their reasoning. John has a false belief about the way Williamson's students formed a certain belief, Jane has a false belief about the possibility of someone truly (and justifiably) believing the negation of Fermat's last theorem.
2) It seems both John and Jane are justified in their false beliefs. Therefore, if they go on to infer that the JTB theory of knowledge is false (assuming that it is false) they will form a justified true belief (effectively, they might be Gettiered, forming a justified true belief which falls short of knowledge that the JTB theory is false). A justified true belief, even if it falls short of knowledge, is epistemically valuable.
3) It is actually controversial, in contemporary epistemology, whether one cannot gain knowledge by inferring from a justified false belief[73]; so it is not uncontroversial that John and Jane could not come to know that the JTB theory is false on the basis of

[73] See e.g. Unger (1968), Klein [1996] and [2008], Hawthorne [2004] (p. 57), Warfield (2005) and Coffman [2008] (pp. 188-194)

those counterexamples involving falsehoods. In particular, it is often held that one can gain knowledge by inferring from a (justified) false premise when, if one were to realize that the premise is false, one could easily replace it with a true (and justified) one. Next point should suggest that John and Jane might meet this condition.

4) Finally, it is clear that John and Jane, supposing they are reasonably smart and willing to spend time thinking about the issue, can easily come to generate more Gettier cases, not involving falsehoods, and thereby come to know (if they didn't already) that the JTB theory is false. If this happens, the initial cases will play a crucial role in the production of such knowledge. The cases will have a cognitive value for John and Jane, allowing them to see the structure of a possible different case.

Given that these points are to be conceded about the two cases I provided, it seems they must be conceded also to the defender of Williamson's account about the Gettier cases in which, according to such an account, our judgement is false. In particular, points 2 to 4 seem to give a rich account of what is still valuable about these 'deviant' cases, which helps to explain our reluctance to think that the deviant realization matters at all. The counterfactual is false, because its antecedent is true and the consequent false, but we have justification to think it is true, and, moreover, we could easily replace the premise with one consisting of a similar, but true, counterfactual. Of course, Ichikawa and Jarvis will still want to insist that, in such cases, there is literally a counterexample, and no false belief is involved. But it is hard to see how this insistence is motivated. At one point, Ichikawa [2009] indicates the disagreement as one about what someone commits herself to when she says, or thinks, " 'the subject has justified true belief but does not know', *i.e.*, when she expresses the Gettier intuition."[74] The foregoing considerations helped to explain some good features of that judgment. But whether the judgment is literally true depends, of course, on what particular Gettier case the person is thinking about. Even on the view defended by Ichikawa and Jarvis, it is not at all trivial to find out exactly what that amounts to. While the cases I provided were constructed in such a way to make it trivial that some false belief is involved, it is not trivial whether some false belief is involved in the case of the stopped watch. So, when someone is shown that the text we are considering is actually realized but it is actually false that the subject is a counterexample to the JTB theory, there should be at least some doubt as to whether the "Gettier intuition" on *that* case was literally true. It is not a straightforward matter.

[74] Ichikawa [2009] p. 438

Let me move to consider briefly the problems for the view defended by Ichikawa and Jarvis (mostly drawing from the discussion in Williamson [2009]). I think the main difficulties for the view can be summarized as follows; the story we build out of the text, in order to make the strict conditional true, has to be extremely rich, for the number of situations which are compatible with the text but do not support the consequent is extremely large. However, having this extremely rich story in the content of the judgment will present a number of drawbacks. Firstly, it will turn out that the cases differ quite a lot between any two subjects considering the thought experiment, since there are many different reasonable but incompatible ways of enriching the scenario. Secondly, even for a single subject, there might be different ways she is disposed to enrich the story, so the content of the case might be indeterminate. Thirdly, the first premise of the argument will become very hard to know, for there might be hidden inconsistencies in the story. All these consequences seem very far from philosophical practice; the advantage of using thought experiments over real cases is often that they present us with short, simple stories, in which all the factors can be surveyed.

Most or all of the difficulties for the suggestion under consideration would be assuaged if we did not require the way we enrich the text to get at the content of the story to sustain a strict conditional. We might somewhat enrich the text, and then use in reasoning a counterfactual premise, saying roughly that if the story were true, then there would be a case of justified true belief which is not knowledge (or what the relevant content is for different cases). So the suggestion that the thought experiment be treated as a piece of fiction and its text supplemented accordingly can actually help Williamson's account to get rid of part of the problem cases. It seems enough has been said by now to warrant taking Williamson's view of the logical structure of the reasoning involved in (at least some) thought experiments as a starting point. In the next section, I will try to respond to a worry for such an account, a worry related to the way we come to know the relevant counterfactuals.

Before getting to that, I still have a preliminary matter to put aside. We have seen that the reasoning involved in the use of thought experiments includes a premise about the possibility of the case. While it is completely uncontroversial that the possibility claim is usually present, it might be expected that I am going to talk about the epistemology of that claim, and in general the epistemology of modality. I am not going to. One might also suspect that I am not giving enough weight to the possibility claim, perhaps misled by the Gettier case example. The possibility claim for a Gettier case is usually trivial, and as we said, there can also be actual Gettier cases. But, one might think, the possibility claim is often the controversial part of a thought experiment, and it is an epistemology of that claim which

would be more interesting. However, that would require an epistemology of modality, and I am not going to present an epistemology of modality in this work. Consequently, I am also neutral on Williamson's project of illuminating the epistemology of modality through the epistemology of counterfactuals[75] (I will come back to this below). To the extent that a satisfying epistemology of modality will require appeal to intuitions, or conceptual analysis, or even a priori justification, the claims I am making on the epistemology of philosophy will have to be weakened. Of course, for the reasons given in previous chapters, I do not believe this will be so. But I am not providing a positive view on the matter.

However, I hope it is clear enough that there is still something interesting to say about thought experiments, even putting aside their modal component. There are, interestingly, other paradigmatic thought experiments that are similar to Gettier cases in having actual counterparts. As the cases I will talk about in the rest of the section are all well-known, I will not describe them, but just mention them. Kripke claims that his Godel-Schmidt case probably has a real life counterpart in Peano and Dedekind. Trolley cases certainly cannot be constructed on purpose, but real life unfortunately provides similar situations. Most cases discussed in the literature on causation are absolutely realistic; it is hard for two people to throw a rock at a bottle so that the bottle would break if either rock hit it, but not *that* hard. Often philosophers look at hypothetical cases not because no actual case would serve their purpose, but simply because it's quicker to present a hypothetical case, or because a hypothetical case might be neater, screening off various kinds of noise coming from the actual ones: think for example of Goldman's 'fake barn county' case or Burge's arthritis case. Many other thought experiments differ in that an actual case is most likely not forthcoming. Even then, the metaphysical possibility of the scenario described by the thought-experiment is often not a matter of dispute. Here is a quick list: Putnam's twin earth (or variants of it); Lehrer's true-temp case; Putnam's brains in a vat; Thompson's violinist; and we could go on. If we put together the cases mentioned, we will find that there is a very vast range of cases from a fairly vast range of philosophical sub-disciplines (epistemology, philosophy of language, metaphysics, moral philosophy) for which the philosophically interesting part of the thought experiment is not the possibility claim, but rather what follows (or does not follow) from it. Of course, there will also be cases in which the possibility of the scenario is seriously disputed (e.g., Chalmers' zombie case, or certain cases in the personal identity literature, and so on). But even those cases will also involve a judgement on the hypothetical situation,

[75] See Roca-Royes [forthcoming] for some criticisms of Williamson's view.

although for those who deem the situation impossible, the judgement will be trivial or anyway uninteresting.

One might think that the fact that the cases involved in philosophical thought experiments are in some sense to be specified out of the ordinary or uncommon makes a difference to their epistemology. I will come back to this in chapter 5. But the point I wish to make here is about the relation between judgements about hypothetical cases and judgements about actual situations which are identical or relevantly similar (where we can find some). Clearly these judgements will have to coincide; if I judge about an hypothetical case that a certain causal relation holds, or that a certain term refers to something, or that a certain action is permissible, than I should give the same judgement about an identically described actual case[76].

It might be useful here to compare the way we reach the conditional judgement involved in hypothetical reasoning with the way the proof of a conditional proceeds in natural deduction[77]. In attempting to prove a conditional, one assumes the antecedent, temporarily treating it as proved, and then one attempts to prove the consequent using the antecedent together with all the resources allowed by the logic (and the previous stages of the overall proof, if there are any). If one succeeds in thus proving the consequent, one has proved the conditional, and the proof of the conditional does not rely on the assumption of the antecedent. The point of the analogy, for my present purposes, is that there is no difference between the sub-proof and a proof which could be conducted if the antecedent had been in fact proved, although of course the latter would then constitute a proof of the consequent and not (just) of the conditional. Similarly, there is a stage of our reasoning about hypothetical scenarios which seems to be independent of whether the scenario is merely hypothetical or rather actual. It is this stage of our reasoning which I will be concerned with.

---

[76] This is actually a problem for the proposal advanced in Malmgren [forthcoming]. Malmgren claims that the content of a judgement about a thought a thought experiment is a possibility claim. We judge that a conjunction is possible: that the scenario obtains and that something is the case in the scenario. But, when presented with an actual Gettier case, we do claim that it is possible that the subject involved does not know. We seem to be justified in the stronger claim that the subject actually has a justified true belief which does not amount to knowledge. This difference is left unexplained by Malmgren's account.

[77] I owe this analogy to Magdalena Balcerak-Jackson.

4.2. A Mysterious Problem

In the next section, I will spell out a possible account of the cognitive capacities we use in evaluating judgements about thought experiments (in particular, counterfactual judgements like 2*; henceforth, I will refer to judgements of that kind, unless otherwise specified), an account which is compatible with everything I said so far. In this section, I will explain why I think, and some other theorists think, that we need such an account, and I will introduce a framework in which the account can be usefully positioned.

A main theme of the present work is anti-exceptionalism, the idea that the cognitive skills we deploy in philosophy, and in the use of thought-experiments in particular, are continuous with cognitive skills we deploy in everyday life and in other sciences (despite their disagreement on other points, Ichikawa and Jarvis [2009] agree on this). Williamson [2007] also insists on this point. He writes that "we assent to [the Gettier verdict] on the basis of an offline application of our ability to classify people around us as knowing various truths or as ignorant of them, and as having various other epistemologically relevant properties"[78]. However, nothing has been said so far on what kind of ability we use when we judge what are the consequences of the obtaining of a certain case, be it actual or hypothetical. I will argue that often what is employed is simply our ability to employ concepts. That we have such an ability, in general, should be not contentious. But one might complain that just talking about our ability to employ concepts in order to explain how we form judgements is not more explanatory than talking about a certain substance's *virtus dormitiva* in order to explain how that substance causes people to fall asleep. Another way to put the worry is to point at the criticism which is sometimes leveled at the postulation of a faculty of rational intuition. The criticism is roughly that we have no idea how such a faculty could fit into a naturalistic picture of the mind, and it is utterly mysterious how it could work; so, even if we currently had no other way to explain how we can gain knowledge in a certain area, making appeal to such a faculty would still have no explanatory value. Why would the same criticism not apply here?

This worry is articulated in some detail in Malmgren [forthcoming]. Malmgren concentrates on the explanatory value of Williamson's proposal about the logical form of thought experiments I discussed in the previous section. Her conclusion is that the proposal that counterfactuals are crucially involved in the logical form of thought experiments does not provide reasons against the view that "intuitive judgments" are a priori. In her words, "the

[78] Williamson [2007] p. 188

weight of the argument against the a priori/a posteriori distinction is carried by considerations that are completely unrelated to (…) the assimilation of intuitive judgements to judgements of counterfactuals" (p. 38). I agree with that conclusion. I presented an argument against the a priori/a posteriori distinction in chapter 3, and it did not rely on the assimilation of "intuitive judgements" to counterfactuals. Nor do I think that Williamson meant any argument he presented against the a priori/a posteriori distinction to rely on that assimilation. However, Malmgren's argument for that conclusion is still interesting. Part of the argument is that Williamson fails to provide a unified epistemology of counterfactuals, and indeed there cannot be a unified epistemology of counterfactuals, and therefore the counterfactual hypothesis has no epistemological role whatsoever.

Williamson sums up his views on the way we form counterfactual judgments as follows:

> one supposes the antecedent and develops the supposition, adding further judgments within the supposition by reasoning, off-line predictive mechanisms and other off-line judgments. The imagining may but need not be perceptual imagining. All of one's background knowledge and beliefs is available from within the scope of the supposition as a description of one's actual circumstances for the purposes of comparison with the counterfactual circumstances (...). Some but not all of one's background knowledge and beliefs is also available within the scope of the supposition as a description of the counterfactual circumstances, according to complex criteria (the problem of cotenability). To a first approximation: one asserts the counterfactual conditional if and only if the development eventually leads one to add the consequent. (Williamson [2007] pp. 152-3)

This passage, as Williamson recognizes, does not amount to a unified epistemology of counterfactuals (he actually writes: "there is no uniform epistemology of counterfactual conditionals"[79]). The reasoning involved in "developing" the supposition can be of different kinds; clearly inductive, as in 'If I were to see a swan, I would see a white swan', or mathematical as in 'If twelve people came to the party, more than eleven people would come to the party'. But, one might complain, this leaves out precisely the interesting cases. We are not told what sort of reasoning is required in developing the suppositions involved in typical philosophical thought experiments. As Malmgren puts the point:

> ..an explanation of our counterfactual judgements in terms of mental simulation does not preclude or make redundant further explanations of some (or all) such judgements and, although

[79] Williamson [2007] p. 152

> it is less permissive than a *virtus dormitiva* explanation, it still does not constrain these explanations enough (Malmgren [forthcoming] p. 34)

Malmgren also makes clear that she does not believe that this is a defect of the particular explanation of counterfactual judgements endorsed by Williamson. In the light of the epistemic diversity we find in the class of counterfactual judgements, she thinks that

> ..*any* hypothesis, *a fortiori* any causal-psychological hypothesis, about counterfactual judgements as such is bound to be very limited in its epistemological implications. We should certainly not expect to get an (acceptable) epistemology for some specific subclass of counterfactual judgements out of it— say, for our judgements of philosophically interesting counterfactuals. (Malmgren [forthcoming] p. 35, original italics)

Again, I do not disagree with anything, and I do not think Williamson would find much to disagree with either. Counterfactuals play an important role in Williamson's philosophy of philosophy, because they provide a solution (or so Williamson claims) to what we might call the "Benacerraf problem" for metaphysical modality, namely the problem of how we are related to modal facts so that we can come to know them. Williamson's answer, in short, is that modal facts are equivalent to counterfactual facts, and everyone concedes human beings can know counterfactuals. I am neutral here on this proposal, as I said above. Counterfactuals are also, as it happens, involved in reasoning about thought experiments, and philosophical thought experiments in particular. But they do not provide, by themselves, an epistemology of thought experiments. One could plug into a counterfactual account of (what we called) the logical form of thought experiments an epistemology for counterfactuals based on intuitions, or analyticity. I take such proposals to have been undermined in the previous chapters.

But we are still left with the question: what can we say about the epistemology of the conditional judgement involved in thought experiments? It will be useful to distinguish different ways to understand the question. On one hand, one might be asking for a general epistemological theory that allows such judgements to be justified, or to constitute knowledge. Not just any theory whatsoever, of course, but the correct theory of what conditions a judgement must meet to receive epistemic appraisal. Developing such a theory is beyond the scope of the present work. But I also think it is not the point of the question. One can fairly easily find a plausible epistemological view that allows for our hypothetical judgements to constitute knowledge. Consider the view defended in Sosa [2007b], on which knowledge (or at least one kind of knowledge) is *apt* belief, and apt belief is defined as a belief that is correct because it manifests an epistemic virtue or competence, which is, roughly, a capacity to discriminate true contents from false ones. The simple view I wish to

adopt is that *a judgement about a hypothetical case constitutes knowledge when it is correct because it manifests a competence in applying the relevant concepts* (keep in mind though that possessing the concepts does not require having a competence in applying them). Call the italicized thesis the Simple-Competence-View (SCV). SCV would differ from Sosa's own view, which is that judgements about philosophical thought experiments, or at least some of them, are of a special kind, because first, they are confined to modally strong contents (necessarily true or necessarily false) and, second, they are explained by a competence that does not rely on perception, memory, introspection, testimony or inference. Call that the Rationalist-Competence-View (RCV). I believe the additional requirements of RCV have to be dropped, for reasons that were explained partly in the previous chapters and to which I will come back in the last section of this one. The judgements on philosophical thought experiments might not be typically based directly on perception, introspection, testimony, episodic memory, or explicit inference, but the competence itself typically is, in my view, epistemically dependent on some of these sources.

However, here I am interested in an objection that applies (although perhaps with different strengths) both to RCV and to SCV. The objection is put forward by Boghossian [2009], and it essentially that "it can look as though we have invoked a mystery to explain a mystery"[80]. In short, the objection is that we have been told nothing about how the relevant competence works. Part of the problem, according to Boghossian, is that Sosa's competence is supposed to work based only on the understanding of the proposition. I agree that this makes the problem more severe for RCV than for SCV, but it seems clear that the objection could be directed to both. The proposal to be developed in the next section should clearly answer Boghossian's challenge, and also any challenge raised by Malmgren.

However, before moving to that, I will consider Sosa's own reply to Boghossian's objection, and explain why I take it to be not entirely satisfactory. This will also constitute one more effort in clarifying the challenge. Sosa [2009], in reply to Boghossian's point, writes:

> ..as I have argued elsewhere, we can appeal to a sort of competence in epistemology even when we have only limited understanding of its modus operandi. This in fact applies not only to rational intuition but also to introspection and even to perception and memory. People could surely know how they knew things even before we gained our vastly improved understanding of how perception and memory actually work. Of course, our knowledge through these various

[80] Boghossian [2009] p. 116

> sources will be enhanced with our improved understanding of their nature and operation. (Sosa [2009 p. 140)

Sosa is certainly correct that we can gain knowledge through the use of a competence even when we do not have a satisfactory account of its working, and examples such as perception and memory illustrate the point vividly. It is overwhelmingly plausible that people with little or no understanding of the workings of such faculties can acquire knowledge through their use, and even come to know that they have acquired knowledge, and, under a certain description, how they acquired it ('how do you know that? I just saw It'), thereby possessing, in Sosa's terms, *reflective* knowledge of the propositions in question. In Sosa's view, reflective knowledge of a proposition requires at least aptly believing that one aptly believes the proposition. But Sosa also claims that reflective knowledge comes in degrees, and the higher degrees "may of course involve scientific and even philosophical perspectives that enable defense of one's first-order belief as apt."[81] This is most likely what he is hinting at where he says that improved understanding of the nature of a competence will "enhance" our knowledge. But, one might say, it is precisely this enhancement which is the object, or one of the main objects, of epistemological theorizing; where else might one look for it? Perhaps, Sosa could say one may look for it not in epistemology, strictly conceived, but rather in the relevant bits of psychology and cognitive science. This is in fact a good place to look. But one need not deny all boundaries between philosophy and empirical disciplines to see the philosophical interest of this investigation. Let us consider again the parallel with perception. Philosophers such as Aristotle, Descartes and Hume were certainly interested in the psychology of perception. Of course, psychology was not at the time an independent discipline. But their interest in this area does not seem, even today, disconnected from their philosophical views. They were interested in providing a coherent and comprehensive world picture, and in particular an account of the place of human minds in the world. Suppose for example that a philosopher defends in a form of Cartesian dualism about the mind. Could she then defend an epistemological theory of perception that says that a perceptual belief is justified when the subject's perceptual capacities give rise to beliefs in a reliable way? Clearly, there would be a problem. How can our perceptual faculties be so reliable? How does the immaterial mind get in touch with the world? In a way, I am attempting to answer a similar question (although hopefully not a similarly desperate one) about our capacity to apply concepts. The question is not purely epistemological, but it is not purely a matter of

[81] Sosa [2009] p. 141

psychology or cognitive science either. It is a problem at the border between these disciplines, or, one might say, a problem at the interface of philosophy and cognitive science.

Finally, I should note that the discussion in the next section will not only prove useful in dispelling the worry that there is anything mysterious about the existence of an ability to apply concepts. Having in mind a possible model of what an ability to apply a concept might be will also be helpful in reinforcing some of the points made in the discussion of the a priori in the previous chapter, and in discussing some epistemological questions to which we will turn to in the last chapter.

4.3. The Application of Concepts

In section 2, I tried to articulate from different angles a certain worry about appealing to our ability to apply concepts; in short, the worry is that any account of our ability to apply concepts which will cover the case of philosophical thought experiments will have to resort to notions such intuition or analyticity. In this section I will look at what psychology and cognitive science can tell us about the way we apply concepts. In particular, I will look at various theories of concepts, and I will argue that this empirical literature provides theories that, although not originally intended this way, can show what an ability to apply a concept is, without appealing to any of the notions mentioned. A few notes of caution are in place. First, there is in this area a vast amount of disagreement. We cannot just consult *the* correct theory of concepts and check what consequences it has for our philosophical concerns. Secondly, sometimes the philosophical interpretation of the theory is equally controversial. Thirdly, I will discuss the theories as if they were incompatible, but many psychologists nowadays prefer a pluralist approach, on which different accounts will hold for different concepts and perhaps even for the same concept[82]. My way of re-interpreting such theories actually makes the pluralistic approach more plausible. Finally, what follows will necessarily be a very rough account of the theories and of their strengths and defects. We could say that we are looking at toy versions of the theories (and toy versions of the objections). But this should still be sufficient for my present purposes. I will proceed by first sketching two different theories of concepts, then I will discuss some objections, and finally I will turn to the connection of all that with my main theme.

Other theories that I might have considered in this connection are the so-called classical theory of concepts, according to which concepts consist of a set of necessary and

[82] See Machery [2009] for an extended treatment on this issue, but see also Rey [2009].

sufficient conditions, and the exemplar theory of concepts, according to which concepts are the storage of a number of exemplars, where each exemplar is identified by a set of properties it possesses. Although some of the details would have to vary (in particular, objections to the classical theory are quite different from objections to the other theories; see e.g. Laurence and Margolis [1999]), the conclusion that I am going to reach later about prototype theory and theory-theory could be defended about these as well. As I am going to explain later, all these theories are implausible as theories of concepts, but they might work as theories of our ability to apply concepts. While the differences matter a lot from the psychological point of view, they are not essential to my present project, so the choice of the two theories I do consider is to a certain degree arbitrary.

The first theory, or maybe better family of theories, of concepts I will look at is *prototype* theories. This kind of approach emerged in the 1970s due especially to the work of Eleanor Rosch (Rosch [1973], [1978], Rosch and Mervis [1975]), as a reaction to the classical theory of concepts, according to which a concept is characterized by a set of necessary and sufficient conditions of applications. The classical theory had come under attack from a variety of fronts; most notably it was thought to be incompatible with the Quinean criticism of the analytic-synthetic distinction. However, the philosophical inspiration for prototype theories was Wittgenstein's notion of a family resemblance concept. The core idea of the prototype theory is that a concept is structured as a set of features that the objects falling under the concept *tend* to have[83]. Crucially, none of the features is necessary. The concept can thus be thought as structured around a prototype, an object possessing all, or as many as possible, of the features. Anything resembling the prototype to a sufficient degree, i.e. having a sufficient number of features, will belong to the category. As a consequence membership comes in degrees. There are various ways of interpreting the latter claim, and they make serious theoretical differences, but the basic thought is that some things are better, or clearer, exemplars of the category than others. To illustrate, consider the concept BIRD. It will consist of a list of features including, let us suppose, "has a beak, flies, has wings, has plumage, lays eggs, build nests, sings, is small (compared to a human)". On this view, a robin is as close to the prototype as possible, having all the features. A chicken is less prototypical, but still a good case. An ostrich is quite far, and a penguin is probably at the far end.

[83] Later versions of the theory gave statistical models which specify numerically the "weight" of the feature in determining membership in the category, which made possible to produce precise empirical predictions. I will concentrate on the simple version of the theory

The theory predicted a series of cognitive effects, which we will call *typicality* effects, that were found to obtain (see e.g. Rosch and Mervis [1975], Laurence and Margolis [1999a]). The main typicality effects are the following. A) Graded membership: people are willing to classify objects in various categories as better or worse exemplars of a category, often with a blurred boundary between very bad cases and things outside the category. This ranking correlates with many cognitive effects, such as b) Retrieval: if asked to name a member of the category, the subjects will name a prototypical member; and if they have to list a number of members, they will do so in order of typicality, and c) Speed and accuracy of categorization: if asked to decide whether an object belongs to the category, the subjects will reach a faster and more often accurate verdict in the case of more prototypical members, and they will give a slower and more often mistaken verdict for less typical members. Moreover, d) Correlation with features: the effects in a), b) and c) can be predicted with an appropriate choice of features possessed by the prototypical members.

I will discuss the problems for the theory after presenting the second kind of theory we are going to look at, since they mostly affect both view, with one exception. The following problem is specific to prototype theories. Armstrong, Gleitman and Gleitman [1983] conducted experiments which showed typicality effects for concepts for which they thought the prototype theory seemed not adequate; in particular they tested the concepts ODD NUMBER, EVEN NUMBER, FEMALE, PLANE GEOMETRICAL FIGURE. For example they found that the subjects were willing to classify 34 as a worse exemplar or even number than 8. But, they thought, it is implausible that the concept EVEN NUMBER has a prototype structure. It seems that something is an even number if only if it is a number and it is even. Various defenders of the prototype theory took the problem very seriously, and they tried to solve it by combining the prototype theory with the classical theory, claiming that concepts are associated both with a prototype and with a set of necessary and sufficient conditions. However, Rosch [1999] takes a different line, and sees it as a "strange twist of logic"[84] that the fact that the theory applies to those concepts as well was used as an argument against it. I think, and here I am anticipating my main conclusion about this debate, that Rosch's reply only makes sense if we understand prototype theory not as a theory of the content of the concepts (for surely 34 is an even number just as much as 8 is, unless we subscribe to a very implausible form of anti-realism about even numbers), but rather as a theory of the way we apply concepts.

[84] Rosch [1999] p. 66

The other kind of theories that I want to consider is *theory-theories* of concepts, i.e. the kind of theory according to which concepts are theories (in a sense to be specified). Cummins [1998], in a paper related to philosophical methodology, describes this kind of view as follows:

> the majority view, I think, is that concepts are theories, either explicit, as in the case of technical scientific or legal concepts, or tacit, as in the case of 'ordinary' concepts

and he goes on to endorse this kind of view (combined with an element of prototype theory)

> My own view, for what it is worth, is that my concept of an elevator is just everything I know about elevators, different bits of which are activated or accessed in different occasions, depending on cues and previous activations, plus some quick and dirty procedures that account for prototype effects. (Cummins [1998] p.121)

There are two important clarifications which have to be added to what Cummins says. Firstly, one could hold a theory-theory view on which the theory constituting the concept is innate, for at least some concepts. But the most common version of this kind of theory holds, quite reasonably, that at least some concepts are acquired, and somehow vary with our knowledge of the world. It is of course a matter of great interest how much of our conceptual apparatus is innate, but I am not presupposing any answer to that question. Secondly, there is much controversy about the nature of conceptual change. In an extreme form of the view, conceptual change is radical; as our concepts change, the new theories are incommensurable to the old ones, in the sense in which according to (a reading of) Kuhn scientific theories are incommensurable to old ones after a shift in the dominant paradigm (on both issues, see Carey [1991] and [2009]).

Margolis and Laurence [1999] cite as strengths of the theory-theory the ability to give a realistic account of categorization judgements and cognitive development. The former is for example found in the possibility for a theory-theory to account for the tendency of adults and, interestingly, children, to take an essentialist view of natural kinds. Contrary to what the prototype theory would plausibly predict, children as well as adults think that external features of a member of a natural kind, such as dogs, are comparatively irrelevant to whether it is a member of the category, and what is relevant is some internal, essential property (Gelman and Wellman [1991]).

It is now time to look at some objections to these theories. There are at least four kinds of problems that are present in the literature (see Armstrong, Gleitman and Gleitman [1983],

Rey [1983], Fodor [1998] and [2008], Laurence and Margolis [1999b]) and that can applied to both theories. They are 1) the problem of shareability of concepts 2) the problem of ignorance 3) the problem of compositionality and 4) The explanatory regress problem. I'll give a very brief sketch of each. Of course the dialectic could be pursued further, with different replies or modifications of the theories for each objection but I will explain later why, problems of space aside, we do not need to do this.

The problem of shareability depends on the fact that prototypes as well as theories associated with a concept (or a word) vary both across subjects and for a single subject over time. However, if this is so, my concept BIRD is not the same as, say, the ornithologist's concept, and the ornithologist's concept is not the same it was when she started her studies (I picked a rather extreme case for vividness of illustration; prototypes will vary across non-experts as well; see Barsalou [1987]; and clearly theories vary across individuals and times, at least if they are understood as Cummins does). This means that it is not clear that our beliefs and the expert's can be in contradiction; which is an unwelcome result, for surely we take the experts to be expert on the very subject matter on which we are not. The problem is particularly severe for the radical version of the theory on which the new concept will often be a theory which is incommensurable with respect to the old one. I must confess it is mysterious to me how it came to be in cognitive science that a theory which makes it completely impossible for people to believe the negation of what they used to believe is supposed to have a strength in its capacity to account for cognitive development.

The problem of ignorance is substantially a form of the problem which was posed by Kripke, Putnam and others against descriptivist theories of names and natural kind terms. We can be radically wrong in our prototypes and theories, and still manage to refer to the relevant objects. If the theories in question wish to give some explanation of the connection between the way they describe the concept and the concept's reference, this seems to be a serious problem; of course they could just not wish to tell a story about it, but then they seem to fail an explanatory burden anyway.

The problem of compositionality has been discussed especially in relation to prototype theories. A complex concept, such as PET FISH, will not have of course the union of the prototypical features of the component concepts (a prototypical fish lives in the sea, or in a river or lake; a prototypical pet can be cuddled). But neither is it simply the intersection; the prototypical pet fish lives in a bowl, but that feature is nowhere to be found for either pets or fishes. So it seems that, on the prototype theory, there just is no way in which the structure of a complex concept can be a function of the structure of the simple concepts on the prototype

theory. This is a huge problem, for it is clear that we can understand a number of new complex concepts far too vast to make it plausible that we are just learning a new independent entry each time. I know of little discussion of this particular problem as applied theory-theories. But the reason does not seem to be that there is some straightforward account of how two theories compose into a complex theory; on the contrary, it seems entirely obvious that there is no mechanical procedure for combining two theories.

The last problem we have to consider is the one I am calling "the explanatory regress". This is the only objection which I am considering that is not, as far as I know, present in the literature already; at least not in this form. Let us ask ourselves, with respect to the prototype theory, how is a feature represented? If a feature is just a property of the members, we would expect the subject to have a concept that applies to that property, otherwise the subject could hardly use the feature to compute the typicality of a member. But, if the concept of the feature is a list of features, and each of those is in a turn a concept, we can easily see that there would be an exponential regress. So the prototype theorist must either maintain that some concepts are not prototypical or that the subjects think of features in some way other than by having concepts. Unsurprisingly, the latter is the alternative prototype theorists prefer. So features could be perceptual traits, recovered as images; but it would seem a naïve form of empiricism to suppose that a concept can be composed of perceptual features *only*. So there must be a way in which information about features is processed without being conceptualized; but it seems that we lack an explanation of how this is done. The problem is even more dramatic for theory-theories. A theory, we would normally think, is made of various integrated beliefs, and beliefs are made of concepts. If each concept is a theory, we seem again to be stuck in a regress. On the other hand, if we think of the information in the concept as somehow non-conceptually organized, talk of "theory" now seems purely metaphorical.

This completes my short survey of the two theories of concepts and of their problems. Of course there are other theories of concepts, some related to these ones, some rather different; the two theories I have discussed here were not chosen because they are particularly plausible. In fact, I must now confess, they were chosen because they are particularly *im*plausible, as theories *of concepts*. My suggestion is that we should think of them as theories of the ability to apply concepts[85]. Interpreted this way, the theories retain all the advantages they had, including the empirical evidence in their favour, for the phenomena they are meant to explain all involve the application of concepts. On the other hand, all the problems

[85] A similar conclusion is reached by Rey [1983] about prototype theories. Rey's considerations clearly apply, *mutatis mutandis*, to theory-theories.

mentioned disappear once the theories are reinterpreted this way. Abilities to apply concepts are not supposed to be shared, and can vary (in fact, hopefully they improve) across cognitive development. They are not supposed to determine the reference of the concept. They are not supposed to compose according to syntactic rules. Moreover, they can consist of associations, either in the form of a prototype or of a tacit theory, between concepts. Finally, we can see how the two theories, as theories of the ability to apply concepts, are not incompatible but rather complementary. We may use prototypes in applying certain concepts and tacit theories in applying others. Even more plausibly, we may use sometimes prototypes and sometimes tacit theories in applying a single concept, depending on the context; for example, in perceptual recognition we might use mainly prototypes, and in inferential processes mainly tacit theories. We may call the combination of a prototype and a theory (and any other psychological structure used in the application of the concept) a 'conception'[86]. Of course, not every conception produces justified beliefs; one can have a very misguided conception of an object or kind of object, thereby being very unreliable in applying the concept. On the other hand, clearly a conception can, in some cases, be sufficiently accurate to be a reliable guide to whether the concept applies. There will be still a further condition that the conception will have to satisfy to count as a competence, to which we will turn in the next section.

Given how easy it turns out to understand these theories as theories of conceptions, it might be thought that what I am suggesting is really superfluous, and the theories are already what I am suggesting they should become, for psychologists really mean 'ability to apply a concept', or 'conception' when they say 'concept', and I am somehow giving them an uncharitable reading. I think there is something to this objection. Ultimately, I would reject such an interpretation of the content of the psychological theories I mentioned. But this is not at all crucial for my purposes. If theories of the ability to apply a concept were already there to be found, this would make my job even easier.

However, I think it is worth pausing on this point. It is worth doing so both because it helps to clarify my stance, and because it is surely worth clarifying some confusion surrounding the concept CONCEPT. One could think that the word 'concept' is really ambiguous, and philosophers and psychologists are often talking past each other. Here are two characterizations of concepts, the first offered by a philosopher and the second by a (philosophically inclined) cognitive scientist:

[86] I take the term from Burge [1991], where a theory of Fregean senses very much in accordance with the present approach is presented.

> Concepts are sub-components of thought contents. Such contents type propositional mental events and abilities that may be common to different thinkers or constant in one thinker over time. Having a concept is just being able to think thoughts that contain the concept. (…) In being components of thought contents, and ways of thinking, concepts are representational or intentional (I make no distinction here). They need not apply to actual objects, but their function is such that they purport to apply; they have intentional or referential functions. (Burge [1993b] pp. 309-310)

> A concept of x is a body of knowledge about x that is stored in long term memory and that is used by default in the processes underlying most, if not all higher cognitive competencies when these processes result in judgements about x. (Machery [2009] p. 12)

Both characterizations are meant to be in some sense fundamental. They should provide one with a preliminary understanding of what a concept is, compatible with many different theoretical views. But they are strikingly different. We may individuate two main differences. Firstly, in Machery's characterization concepts consist of "knowledge". Even if here the term knowledge is used in a deflationary way, as roughly synonymous with "true belief", Burge's concepts do not consist of "knowledge". I take knowledge to require a complete thought, but concepts for Burge are only sub-components of thoughts. Secondly, Burge's concepts aim to apply or refer. There is no mention of such categories in Machery's characterization. The concept (the body of knowledge) is used in reaching a judgement, but it is not even clear if it would make sense to say of a body of knowledge that it literally "refers" to an object.

For the reasons explained above, I take concepts to be what Burge says they are[87], and I take Machery to be wrong. But Machery's characterization would be roughly correct if he were talking about conceptions. Does this provide a sufficient reason to say that psychological theories are really about conceptions, or ways in which we apply concepts? There is a sense in which what psychologists are really interested in is indeed the ability to apply concepts. However, this does not warrant by itself the suggested semantic reinterpretation. There is an alternative way of describing the situation, which is at least equally natural, and that is to say that often psychologists mistakenly identify the ability to apply a concept with the concept itself. Their theories are in a sense about conceptions, but they contain the mistaken assumption that a concept just is a conception. Here is an analogy: Columbus, the explorer, could say a few true and interesting (to his contemporary Europeans at least) things about America, but he did not realize that it was America he was talking about. Suppose he called the place he had reached 'China'. This description of his situation does not warrant

[87] Of course that's not a *theory* of concepts. I do not have one, much less an original one. Various theories of concepts are going to be compatible with that initial characterization.

interpreting the word ‘China’ as meaning what we mean by ‘America’. In fact, given that he had wrong beliefs about the size of the earth, it is much more natural to say that he was convinced to have reached China. Again, nothing I say here hangs on assimilating psychologists to the confused Columbus. But nothing excludes that they are confused, and that they provide us with some rudimentary maps of the ability to apply a concept, even though they take themselves to be investigating concepts *simpliciter*.

My aim, let us recall, was to point at what form the ability to apply a concept could take, such that it is compatible with what we currently know about the mind, and it could deliver judgements about an hypothetical case. Clearly, the theories that I have described respect the first requirement. How about the second? It seems that they fully respect it. Take the paradigmatic case of thought experiment provided by Gettier cases, e.g. the one I provided at the beginning. Our prototype or implicit theory of knowledge presumably includes some kind reliable connection between the subject’s belief and the truth of the proposition involved. Perception provides one such clear case. The lack of such feature in the belief involved in the Gettier case yields a negative verdict; the lucky belief is not knowledge. On both the theories considered, there is no need for the subjects to be explicitly aware of what is involved in their capacity to apply a given concept. Thought experiments therefore can serve the purpose of making such information explicit. This is also, of course, reminiscent of what Sorensen [1992] calls the recollection model of armchair inquiry, a model going back at least to Plato. On this model, in considering new cases, actual or hypothetical, we tease out our own view of the subject-matter involved, a view we already possess but need to make explicit.

The theories I described meet fully at least part of the worries highlighted in section 4.2, in particular Boghossian’s worry that we were appealing to a mystery to explain a mystery. On the contrary, we are appealing to relatively well-understood mechanisms to explain a mystery. We are not appealing to a *virtus cognitiva*, in appealing to the ability concepts. We are appealing to psychologically plausible mechanisms, and we have not of course appealed anywhere to analytic connections or intuitions in the sense of a separate mental state. In the next section, I will say something more about the consequences of the proposal, in particular with respect to the epistemic properties of the judgements.

4.4. Learning to Apply Concepts

In this section, I will further explain some aspects of the view I am proposing. I will start by considering an objection, which will lead me to clarify some features of the view. Then I will end by comparing my view to a similar one advanced by David Papineau.

Williamson objected to a previous version of this proposal, one restricted to prototypes, that although it seems plausible that the prototype plays a role in the ability to apply a concept, the latter cannot be identified with the prototype, for two people could have the same prototype but different abilities to judge how similar something is to the prototype. I think the point can be fully accommodated by two kinds of consideration. First, we have to note that there is a standard competence-performance distinction. Certainly two subjects might have the same ability to apply a concept, while one of them reaches more correct judgements[88]. For instance one of the subjects might be drunk, or generally unreflective and overconfident, and so on. Once all such factors affecting performance are ruled out, my suggestion is that there would be no residual difference in the judgements. The prototype is not (just) an image that we have in the mind, which we have to compare to different objects. To associate a certain prototype with a concept requires being disposed to judge accordingly, once performance errors are ruled out. On the view I am suggesting, the competence in applying the concept simply includes the competence in judging how similar something is to the prototype. Similarly for having a tacit theory. Having a tacit theory will entail being disposed to judge according to it. For any way of specifying any ability, one can always in principle ask about our ability to make use of the ability (the competence in moving from competence to performance); I see no particular worry about the specific form of the suggested account.

However, there is a further respect in which two subjects who share a prototype can differ, and that is the strictly epistemic respect. While two subjects who have the same prototype and perform equally well will reach the same judgments, the epistemic status of those judgements might differ. The point is more easily illustrated by considering the ability to apply a concept understood as an implicit theory. Our theories are not typically innate. Old beliefs can be abandoned and new ones can be added, responding to experience and general revision processes. Now, this means that there is a sense in which the theory constituting the ability to apply a concept might well be unjustified. The beliefs that constitute the ability might fail to have been formed as the result of an appropriate response to the empirical

[88] A further reason why two subjects with the same ability to apply a concept might reach a different number of correct beliefs is that those believe are formed in ways that involve directly some other belief-forming method. I take the ability to apply a concept to be involved in all belief-forming processes, but usually, or at least often, our beliefs also depend epistemically, in a direct way, from some other source, such as perception, testimony or episodic memory. In those cases, the reliability of the way we apply concepts has to be judged conditionally on the correctness of the input they take.

evidence and other rational considerations. So the following is a genuine possibility: that two subjects have the same implicit theory, and perform equally well in applying it, thereby being equally reliable in getting the extension of the relevant concept right, but for one of them the resulting categorization beliefs constitute knowledge, and for the other they do not, because they are driven by beliefs that are not themselves justified. The same point, I believe, applies to prototypes. It is uncontroversial that prototypes associated with a concept can change across cultures, across individuals, and across different times for even a single individual. Of course if we were to identify the prototype with the concept, this would mean that we would have different concepts, but since we are now thinking of the prototype as an ability to apply a concept, this just means that the ability to apply the concept changes. Although it is not clear whether we can say that the prototype itself is "justified", for it does not consist of beliefs, we can surely say that one prototype is more reliable than another, and more importantly for my present point, that some prototypes are formed in epistemically better ways than others, e.g. responding appropriately to interaction with the evidence which the subject had. To sum up, in order for the conception to constitute a competence in a normative sense, it is not sufficient that it is reliable, in the sense of providing a sufficiently high ratio of correct judgements. The way the conception was acquired is also essential to its epistemic effects[89]. This means that a strong form of epistemic externalism holds here; two individuals internally identical, who possess the same concepts and inhabit at present the same environment, may differ in the epistemic properties that their reasoning instantiates, because of a difference in their causal history.[90]

There is a consequence of this picture of the conceptions underlying a competence in applying a concept, and of the way they are acquired and modified, which is of great importance for the self-understanding of philosophy. On this picture, a judgement we reach about a thought-experiment by making use of our competence in applying a concept, or even making use exclusively of that competence, need not be a priori or analytic. The competence required to reach the judgment will typically exceed what is required for concept possession, and it will be epistemically dependent on past experiences. So this confirms the conclusion I had reached in chapter 3.

[89] This thought might be spelled out in modal terms; if a conception was acquired improperly, there seems to be a sense in which the subject employing it could have easily formed a false belief.

[90] I do not rule out the existence of interesting epistemic properties that supervene on the totality of the subject's present mental states, but I do not see them as relevant in this context.

Papineau [2009] also argues for the view that judgements about philosophical thought experiments are the product of a capacity to apply concepts, and that such capacity encapsulates (to use his term) empirical information. I am of course also defending this general claim. However, there are some rather important differences between Papineau's account and mine (besides the fact that my account goes a little further than his in providing an account of the capacity itself), which it will be useful to look at.

Papineau thinks that "..substantial synthetic assumptions are built into the automatic mechanisms that allow us to make particular judgements about philosophically salient categories like knowledge, names, persons, free will and so on."[91] He also argues that this supports the view that the relevant judgements are a posteriori justified. As I explained in the previous chapter, I am sceptical about the usefulness of the category of a posteriori justification, just as much as I am skeptical about the usefulness of the category of a priori justification. But I am in agreement with Papineau that experience plays some role in shaping our capacity to apply concepts. However, I am in disagreement with him where he talks about an "automatic mechanism". Let me clarify. Papineau thinks the relevant capacity always operates at a sub-personal level. He compares the working of the relevant capacity to the way in which the visual system computes the shape of a represented object starting from sharp changes in intensity in the stimulation of the retina[92]. I have talked about 'implicit' or 'tacit' theories. But this is not the same as situating the theories at a sub-personal level. A belief that most readers will share is that they are not giant pink giraffes. This is a tacit belief, but it is held at the personal level, and it is easily retrievable (different tacit beliefs might be less easily retrievable). By contrast, beliefs about the relation between the stimulation of the retina and the objects we are seeing are not retrievable, and they are not attributable to the subject. I do not want to rule out the possibility that some automatic, sub-personal mechanism, is involved in the capacity to apply concepts. But I do want to exclude that *only* automatic, sub-personal mechanisms are involved. On my view, the mechanisms involved may range from completely sub-personal to completely conscious, with all the intermediate degrees.

Papineau believes that describing the ability to apply concepts as an automatic mechanism helps in explaining the appearance that judgements about philosophical thought experiments are not falsifiable. I do not believe there is any such appearance, at least not a strong one. Judgements about thought experiments are often uncertain. The "sub-personal

[91] Papineau [2009] p. 18

[92] Papineau [2009] p. 17

picture" of the capacity to apply concepts leads Papineau to a moderately pessimistic view of the epistemology of judgements about cases. Later on in the paper he writes:

> The function of cognitive mechanisms that embody encapsulated assumptions is to deliver judgements about particular cases quickly and efficiently. Because of this, the relevant assumptions are standardly rules of thumb that work well enough in most cases but are not strictly accurate, in the way illustrated by the familiar perceptual examples. If the cognitive mechanisms behind philosophical intuitions are at all similar, we should expect encapsulated philosophical assumptions to have a similar status. They may work well enough for practical purposes, but they may not be strictly accurate and may lead us astray in certain cases. If we are to be confident about these assumptions, we will need to make them explicit and subject them to proper a posteriori evaluation. (Papineau [2009] p. 21).

As I said above, I do not find Papineau's motivations for his description of the capacity underlying hypothetical judgements very convincing. The resulting form of moderate scepticism is also worrying (I will argue that my view does not lead to similar results in the next chapter). More importantly, Papineaus's picture is in sharp contrast with recent developments in psychological research.

A kind of view which enjoys growing popularity in the psychology of reasoning is represented by "dual-process theories". According to this kind of view, humans possess two reasoning systems, often labelled system 1 and system 2 (see e.g. Frankish and Evans [2009]). System 1 is, very roughly, evolutionary old, unconscious, automatic, fast, based on associations. System 2 has the opposite features: evolutionary recent, conscious, voluntarily controlled, slow, based on logical reasoning. At first glance it might seem that this conforms very well with Papineau's view. He is claiming that "intuitive" judgements proceed from system 1 reasoning. However, by proponents of dual process theories clearly hold that hypothetical reasoning triggers system 2 reasoning, the kind that is exactly opposite (Evans [2007] offers a detailed account of this claim, as well as much empirical support).

I believe that, if one considers the matter carefully, the view that philosophers form judgements about thought experiments *always* in a way which is *completely* inaccessible to conscious reflection ought to strike one as implausible also on purely common-sensical grounds. In addition, for those who are philosophers, Papineau's claim should seem implausible on purely introspective grounds. Philosophers spend long amounts of time in careful reflection about hypothetical cases. We should not assume that this is a waste of time without good reason. The more substantial philosophical point is that in considering a hypothetical scenario we allow ourselves to form judgements on the basis of all of our

background knowledge, implicit and explicit, and whatever its origin, unless it contrasts with the assumption that the scenario holds.

Conclusion

I have been arguing that we can gain knowledge through the use of thought experiments by employing no special faculty and no capacity essentially different from the ones required to yield judgements about actual cases. In particular, I have argued that we can gain such knowledge through our ability to apply concepts, and that some psychological theories of concepts, the prototype theory and the theory-theory, can be reinterpreted as providing a model of such an ability.

Chapter 5
Experiments, Scenarios and Experts

Beginners are as unskilled at thought experiment as they are at ordinary experiment. They do not know what to look for, they don't know how to get it, and they don't know how to report what they do get.

Roy Sorensen, 1992

Introduction

In chapter 4 I have defended a view of thought experiments. On that view, thought experiments are just instances of ordinary hypothetical thinking. In philosophy, we often make judgements about thought experiments; such judgements have counterfactual form and are the product of our ability to apply a concept – an ability which is distinct from possession of the concept itself, and empirically shaped. However, I noted that I had not yet addressed a certain kind of epistemic worry, the worry that the sort of hypothetical scenarios considered in philosophy are too far-fetched for our ordinary capacity to apply concepts to give a verdict about them in a reliable way. The main topic of this chapter is a kind of empirically motivated challenge to the use of thought experiments, which may also be seen as a systematic development of that worry. The main idea of this challenge is to investigate empirically what judgements people tend to give about the sort of cases that philosophers consider, and using statistical methods that are standards in psychology to establish by what factors these judgements are influenced. Since the start (e.g. Weinberg et al. [2001]) this kind of investigation has been conducted in the light of the hypothesis that judgements about philosophical thought experiments might be influenced by factors which are irrelevant for their truth, such as the social or cultural background of the person who makes the judgement. Much discussion ensued, which led proponent of the challenge to clarify and sharpen their

arguments, and to conduct much additional research. This process is still going on. However, I will argue that, at this stage, very little has been established by the proponents of the challenge.

In the first section, I will describe carefully the form of the challenge, and one use of thought experiments which I want to defend from the challenge. The second section will develop a first kind of reply. Given that thought experiments involve, in my view, our ordinary capacity to apply concepts, successfully challenging thought experiments in general would mean establishing an extreme form of scepticism. The proponents of the challenge need to develop the idea, mentioned above, that there is something peculiar about philosophical thought experiments; but I will argue that they have not done so, and that an initially promising way of developing that thought (building on the view I defend in chapter 4) does not yield interesting epistemological results. In the third section, I will defend a further reply to the challenge, based on the idea that philosophers may possess some forms of expertise in evaluating thought experiments which non-philosophers lack.

5.1. Thought Experiments and the Experimentalist Challenge

There are various ways thought experiments might be relevant to philosophical theorizing, even limiting ourselves to their use in building arguments, as opposed to raising interesting questions. It will be useful to distinguish them clearly, to understand the problem I am going to discuss later. One might take as a premise the claim that most people give a certain judgement about a certain kind of thought experiment (hypothetical case), and proceed to use this premise, perhaps as the basis of an inference to the best explanation, to draw further conclusions, e.g. about philosophically relevant aspects of human psychology. If this is what one is doing, then obviously one should have adequate justification for the premise one is using, the claim that most people give that certain judgement about that kind of hypothetical scenario. It is generally not good enough to give one's own judgement on the hypothetical case and assume that most people will give the same judgement. A theorist must at least undertake some informal inquiry to get a sense of how typical her judgement is in that respect, such as, for example, comparing her judgement with those of her colleagues or other competent speakers. Sometimes, this might be enough. Indeed, Jackson [1998] briefly argues that this is usually enough: "Everyone who presents the Gettier cases to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases. But it is also true that often we know that our own case is typical and so can generalize

from it to others"[93]. In the first decade of the new century, a movement of thought has developed that advocates, among other things[94], a more thorough empirical study of such assumptions. This movement is usually known as the 'experimental philosophy' movement, or the 'X-phi' movement. Initial results (Weinberg et al. [2001], Machery et al. [2004], Swain et al. [2008]) seem to show that Jackson was overly optimistic. Philosophers are less typical than they thought, and probably Jackson did not know what answer people get in presenting students with Gettier case, as he was wrongly assuming they get (what according to most of the philosophical community is) the right results, while the truth is more complicated. Be that as it may, this is not the sort of use of thought experiments I am concerned with here. I am interested in a more direct use of thought experiments, the use of the hypothetical judgment itself about a case, and not of the judgment that we make or are inclined to make the judgment. Typically, the judgment I am thinking about is of the form 'concept C applies applied to X in scenario S' and not 'I (we) am (are) inclined to apply concept C to X in scenario S'.[95] I am not taking a stand on how prominent these two different uses of thought experiments are or should be in philosophy, I am willing to be pluralist about both issues, at least for the sake of the argument.[96]

The main claim I am defending in this chapter is that a subject can often form justified beliefs, which will constitute knowledge if everything else is fine, about hypothetical scenarios of the kind typically used in philosophy, even taking for granted the empirical results obtained so far by experimental philosophers, and taking for granted that the subject knows about those results. A presupposition I will be using here is that the judgments we make about hypothetical cases have a definite answer. One might worry that in some of the cases philosophers consider there is no determinate answer as to whether a certain concept applies, but there are incompatible and equally acceptable ways of extending our usage with

[93] See Jackson [1998] pp. 36-7.

[94] For a survey of those other things, see Knobe and Nichols [2008]

[95] One might also reason as follows: 'Most people are inclined to apply concept C to X in scenario S', therefore 'Concept C applies to X in scenario S'. At this point, I am also not interested in this kind of use of thought experiments. It makes a difference if the form of argument used by a philosophers is instead: 'I am inclined to apply concept C to X in scenario S', therefore 'Concept C applies to X in scenario S'. My defense of the use of thought experiments would also apply to this first-person style of argument (which was however criticized in chapter 1).

[96] But see Cappelen [forthcoming] for a strong case that in actual philosophical practice premises about the philosophers mental states are not typically used.

respect to such cases[97]. While I think such a situation is possible, and probably even actual, I am going to presuppose, at least at this stage, that it is sufficiently rare to be ignored, and that in the cases we are going to discuss there is a fact of the matter as to the application of the concept in the relevant hypothetical case. I will come back to the significance and the motivations of this presupposition below. A further presupposition will be that there often is genuine agreement and disagreement (not merely verbal), both between philosophers and between philosophers and non-philosophers, about the correct judgements to be given on a thought experiment. Other things being equal, disagreement on a sentence or utterance between two parties might be evidence that the sentence or utterance is used with the different meaning, so that the disagreement is merely verbal; but only very weak evidence. Again, I am not saying that verbal disagreements never take place, but I am going to assume that they are sufficiently rare to be ignored for my present purposes. Note how the second assumption, that disagreement about thought experiments typically does not indicate difference in meaning, is favorable to the experimental philosopher, or in general to someone who wishes to put in doubt the reliability of judgements about thought experiments. Sosa (Sosa [2007c], [2010]) argued convincingly that if there is a divergence in meaning between philosophers and non-experts subjects, the experimentalist challenge is undermined. I accept that conditional. Sosa also argues that the antecedent of the conditional is true. I am not entirely convinced by those arguments on a general level, but I am also not going to try to respond to them on behalf of the proponents of the challenge.

Experimental philosophers take their results to challenge the main claim I am defending here. They believe that the results obtained so far already show that the practice of appealing to thought experiments, at least in its current form, is seriously compromised. This challenge is articulated by Weinberg et al. [2010] in the form of an argument; here is how they describe it.

> First, there are the experimental results themselves, which at this point in time overwhelmingly concern 'ordinary' subjects (typically, but not entirely, university undergraduates). In particular, there is the claim that the studies in question reveal a worrisome pattern of responses in those subjects, such as ethnic variation or sensitivity to the order of presentation of the cases. Second, there is a metaphilosophical claim that the relevant philosophical facts do not pattern in the same way; there is a misalignment between the contours of the philosophical facts and the contours of philosophical judgments (or, at least, of those judgments as studied by the experimental philosophers). Third, there is an ampliative inference

[97] Unger [1984] defends this view. I will come back to its difference with the experimental philosophy challenge.

> from the patterns disclosed concerning the ordinary subjects to the predicted occurrence of those patterns in professional philosophers. Putting all three pieces together, the armchair practice is thus challenged: we have reason to think that it involves deploying a source of putative evidence that is sensitive to non-truth-tracking factors. (Weinberg et al. [2009] p. 332)

I find this reconstruction of the experimental philosopher's critique rather useful. We can put it in a clearer argument form as follows. First there is an argument relative to non-philosophers

1) Non-philosophers' judgments about cases of class C are sensitive to factor F.
2) Factor F is irrelevant to the truth of the judgments of class C.
3) If 1 and 2, then non-philosophers' judgments about cases of class C are epistemically defective.

Hence, non-philosophers' judgments about cases of class C are epistemically defective (from 1, 2 and 3)

But the argument can also continue after premise 2 as follows

3*) If 1, then philosophers' judgments about cases of class C are sensitive to factor F.
4) Philosophers' judgments about cases of class C are sensitive to factor F (from 1 and 3).
5) If 2 and 4 then philosophers' judgments about cases of class C are epistemically defective.

Hence, philosophers' judgments about cases of class C are epistemically defective (from 2, 4 and 5)

The argument I am interested in here is of course the latter (1, 2 ,3*, 4, 5 – call it the Challenge Argument). To be more precise, what we have here is not an argument, but rather an argument schema. At least on some ways to fill out the schema with respect to C, I am going to grant premise 1 and, with some reservations we will come to later, premise 2, two of the "three moving pieces" Weinberg et al. are talking about. My reconstruction brings out a complexity in the second moving piece; the claim is not just that the factors the relevant judgments are sensitive to are irrelevant to the truth of such judgments, but also that this is epistemically relevant and it represents a serious kind of defect (premise 5). I think this is true. However, it is worth noting that there might be a problem of self-defeat lurking here. According to this line of thought, if the argument were successful, then we would lose our

justification for premise 5. I will not discuss this line of objection here, although I will consider a different worry about self-defeat.

I will consider two lines of reply to the Challenge Argument, both of which I regard as sufficient to undermine it. The first line of reply starts with a request for clarification; what is the class of judgments we are talking about? On some specifications of the class, the argument will turn out to be sound, but uninteresting. On other, more ambitious specifications of the target, the argument will turn out to 'prove too much', leading to undesirable forms of skepticism about hypothetical judgment, which suggests that one of the premises must be false. I will call this the target problem.

The second line of reply will challenge the plausibility of premise 3*). It is worth noting that this ampliative step in the argument had gone unnoticed until it was pointed out by responses to the challenge, especially in Ludwig [2007] and Williamson [2007]. My guess would be that the reason why the step went unnoticed is that it is unnecessary when we criticize some of the different uses of thought experiments I was talking about above, in which philosophers directly appeal to what non-philosophers would say. Be that as it may, Weinberg et al., having realized the problem, attempt to put a patch on it by suggesting that the psychological literature on expertise makes it implausible that philosophers possess some kind of expertise in assessing thought experiments. However, I will argue that firstly, the literature they cite does not really support their claim, and, secondly, there is a part of the psychological literature on expertise which they fail to consider and which clearly tells against premise 3*.

There is a different way in which one could solve the problem for experimental philosophers, and that is conducting empirical studies that directly support premise 4. These empirical studies should involve philosophers as subjects. However, at present few such studies have been conducted, and they only concern one area, moral philosophy (Schwitzgebel and Cushman [forthcoming], Schulz et al. [forthcoming])[98]. So the existing

[98] A possible exception has been pointed out to me by Joshua Knobe. Sytsma and Machery [2010] find some differences in philosophers and non-philosophers judgements about a case in which a robot discriminates a red object from objects of different colors; it turned out that non-philosophers where much more prone to say that the robot "sees red. However, the case seems to me under-described, and it is unclear in any case what the right judgment would be (if there is one), so the study cannot tell against the expertise hypothesis. Sytsma and Machery conclude from their study that philosophers and non-philosophers are using different concepts. If that is the case, again, there is nothing here that helps the Challenge Argument. However, the difference in the concepts employed does not follow from the difference in judgment, so Systma and Machery seem to be using an invalid argument. Interestingly, they use "concept" and "conception" interchangeably, which hints at a possible

studies might at best provide grounds to eliminate the ampliative step in an argument for the limited conclusion that philosophers should not appeal to thought experiments in that area. I will come back to this in the next section. More importantly, there are some methodological difficulties in constructing such studies which seem hard to solve. Some, perhaps not insurmountable, difficulties have to do with the choice of a scenario to ask subjects about. The scenarios used in most X-phi studies were thought experiments, sometimes really famous ones, taken from the philosophical literature, or easily recognizable variants of those (see e.g. Weinberg et al. [2001], Machery et al. [2004], Swain et al. [2008]). A professional philosopher will usually have a previous opinion on those, which will be strongly influenced by a number of factors, including at least education and peer-pressure. Therefore the significance of the test would be diminished. On the other hand, using scenarios which have been constructed for the purpose of the experiment would also reduce the significance of the result. Every case is different, and each case might present peculiar difficulties. Ideally, in the experimental design one should use a case which is analogous in significant ways to well-known philosophical thought experiments, but not *obviously* analogous to the well-known cases; of course, this is not a trivial task. Even more serious difficulties have to do however with the choice of the subjects to be involved. It is highly unclear to me how we could individuate 'the experts' in a sufficiently neutral way. First of all, philosophy, at least as practiced today in the analytic tradition, is divided in a vast number of specialized sub-fields. I don't think this can be ignored, if we want the experimental data to have some bearing on the way philosophy is actually practiced. If one has never published a paper in metaphysics, her judgment on a thought experiment in that field will have very little impact to the research community; if you have never published in a subfield, you are very unlikely even to referee a paper in that subfield. Moreover, we would also have to choose a cut-off point on the degree of professional success required. That is not trivial. It is generally acknowledged that social recognition is not a perfect guide to expertise. For example, Schwitzgebel and Cushman [forthcoming] use possession of a PhD as the criterion of expertise. Since possessing a PhD is a necessary but, unfortunately, not a sufficient condition to be a professional philosopher, this is clearly not a good criterion. It might be that not even being a professional philosopher working in a certain area is sufficient to be an expert in the relevant sense. It might be that among, for example, metaphysicians, only 60% possess the kind of ability in judging thought experiments that really separates them from the non-experts. If that were the case, of course

explanation of their hasty conclusion. I argued in chapter 4 that some theorists systematically confuse concepts and the capacity to apply concepts; it might be that this is such a case.

we could not individuate the experts simply on the basis of professional success or some other more or less measurable device (I make the metaphysically controversial assumption here that we do not live in a perfect world – besides, other metaphysicians might be more proficient in some different way). Maybe there are ways around this difficulty as well, and I will even have a suggestion to advance for those interested in doing experimental studies on philosophers. However, at present it seems justified to look for indirect ways of evaluating the plausibility of the ampliative step in the Challenge Argument. We will do that in section 5.3; first, we will need to give a closer look to the target of the argument.

5.2. The Target Problem

Weinberg, in a paper aptly called "How to challenge intuitions empirically without risking scepticism" (Weinberg [2007]), addresses the worry which I am going to discuss in this section, in the form in which it was presented in Williamson [2004]. The worry can be described starting from the observation that early X-phi literature used the term 'intuition' without providing a psychological or philosophical theory of what intuitions are. Williamson argues that, as it is used in the philosophical community, the term can refer to any sort of judgment or belief. People involved in the X-phi movement (unlike, for example, someone like Bealer, as we've seen in chapter 1) do not have strong theoretical reasons to disagree with this view. But then the claim that intuitions are unreliable, or that they are not justified, or anything in the ballpark, turns out to lead to an incredibly radical form of scepticism, scepticism about human judgment in general. Aside from being, to say the least, an undesirable result, this scepticism would have to apply to the judgment involved in the arguments presented by the experimental philosophers as well, so the position would end up being self-refuting. Limiting the attention to hypothetical judgments would scarcely help; for one thing, this limitation would seem unprincipled. For another, it would still result in a form of scepticism about a large part of human cognition. Moreover, the overall strategy would still seem to be self-refuting; it is very hard to reason without conditionals, and any conditional seems to be equivalent to the assertion that in the hypothesis that the antecedent is true, the consequent also is.

One reaction to the target problem is to limit the relevant class of judgments under attack by concentrating our attention on the subject matter. For example, I mentioned in the previous section that some studies concentrate on thought experiments that have to do with moral judgements. Now, the proponents of these studies are not quite explicit about the consequences of banning the use of hypothetical reasoning in the moral field. It seems

obvious to me that this would require a fairly radical moral scepticism, or relativism. If we are able to objectively judge that any action, such as for example a specific act of torture, is wrong, then presumably we were previously able to judge hypothetically, with the same degree of objectivity, that if one had done that action in those circumstances, that action would have been wrong. I find this kind of view unattractive. But it is not self-defeating, as instead a similar scepticism about epistemic evaluation would be[99]. I am not going to argue in this section against scepticism restricted to the moral domain.

But the line of response to the target-problem defended in Weinberg [2007] is not to limit the relevance of the argument to one area. The response is rather that, even granting everything I said so far, it does not follow that hypothetical cases of the kind commonly used in philosophy under the label 'thought experiments' do not pose special problems. The reason, according to this line of thought, is that very often the cases we considered in philosophy, although they might well be clearly possible, are out of the ordinary or uncommon; or, as Weinberg puts it, they are often "esoteric, unusual, far-fetched or generically out-landish"[100]. For short, I will use the technical term "funkiness" to indicate the property of philosophical thought experiments that is supposed to make them epistemologically both distinctive and distinctively problematic.

A first problem for this idea is that it's clear that the pre-theoretical conception of something being unusual, far-fetched or out of the ordinary cannot be the notion employed here. Consider the following situation: a pink elephant stands on the back of another pink elephant on an airplane parked in the middle of the street in a big city, and a third pink elephant climbs successfully on top of the first two. This is an uncommon situation by everyday standards, but to give judgements about what happens in the case does not seem at all hard. There would be three pink elephants standing one upon the other on top of the airplane in the middle of the street[101].

A fairly obvious suggestion to explicate funkiness would be using some criterion of modal distance, so that the more a possible situation is dissimilar (in the relevant way) from the actual world, the more funky it is. On this reading, the experimentalist objections could perhaps be buttressed by evolutionary considerations. Reliably applying concepts to actual situations and situations which could *easily* be actual will be evolutionary advantageous, but

[99] Moreover, there are meta-ethical views, such as most forms of error theory, that entail this form of skepticism.

[100] Weinberg [2007] p. 320

[101] Cappelen [forthcoming] makes a similar point.

being able to apply them to distant possibilities will not do much. However, this proposal will not serve the purpose of criticizing philosophical thought experiments, for, as we've seen at the beginning of section 2, many 'thought experiments' concern actual cases, or could concern actual cases if we thought this would somehow improve on their standing. Moreover, the example used above could also make the point that it is in fact far from obvious that modal distance per se will create epistemological problems.

I will here do some work for experimental philosophers, trying to develop a different and, I think, more interesting suggestion about what funkiness is. This suggestion would involve the very same theoretical machinery I have deployed in chapter 4 in exploring the question of what an ability to apply a concept might be. A funky case will present us with an object which, talking in terms of similarity to a prototype, will be far from being clearly similar enough to the prototype to count as a member of the category, but also not dissimilar enough to be clearly not a member of the category. Or, talking in terms of theories, the object will not satisfy enough of our beliefs to be counted by the theory as clearly belonging to the relevant category, but not quite so few to be counted as being clearly outside the category. We could call this kind of cases 'classificatory dilemmas', or, more imaginatively, 'platypus cases', honouring of course this animal which presented biologists with a similar problem[102]. Putting together the two bits of technical terminology I introduced, the experimentalist objection would be that philosophical thought experiments are funky because they are platypus cases; this explains why we are not reliable in giving judgments about philosophical thought experiments. It is worth observing that such an objection does not really need to rest on the specific empirical findings. It could be brought out on quite general grounds. In fact, a *prima facie* very similar objection to the use of thought experiments has been laid out, at least with respect to thought experiments in a certain specific sub-field of metaphysics, that of the debate on the notion of person, by Tamar Szabó Gendler (Szabó Gendler [1998], [2002]).

I still see two problems for this understanding of the target of the X-phi critique. The first problem is that it will not be general enough to make thought experiments used in philosophy a worrying category, since not all philosophical thought experiments are platypus cases and not all platypus cases are problematic. The second reply is that even if we found a significant number of cases that are hard in the way platypus cases are supposed to be, it

[102] I want to avoid instead talk of 'border-line cases'; such an expression suggests cases in which a concept has roughly a single gradable criterion of application, and the object is between the points on this single scale which clearly warrant the application or the non-application of the concept; a paradigmatic platypus case, on the other hand, is one in which different criteria for the application of the concept conflict.

would not be clear what interesting epistemological consequences this would have. I will now look at the two replies in more detail.

Let me start the first line of reply with a dilemma. Do we have some independent theoretical way to individuate this kind of cases, or do we individuate them by first noting that they present us with a difficult problem? The option of embracing first horn of the dilemma seems precluded, because I just don't know of any such independent theoretical criterion. We have no way of measuring the similarity to the prototype, or the amount to which a tacit theory should be correct, which are required for the application of the concept. The second option leaves us without a way to motivate the claim that philosophical cases are hard; the claim that they are platypus cases was supposed to do that, but it is itself, it turned out, only motivated by the claim that philosophical cases are hard. So we have no principled reason to think that all philosophical thought experiments are hard. But maybe we have inductive reasons, so to speak, to think so? Not at all. Many philosophical thought experiments are easy. First, there are cases which get little attention, but still play a role in philosophical theorizing. We know (moral scepticism aside) that torturing an animal for fun is wrong, and that a completely unjustified and unreliably formed true belief is not knowledge. We can come to know these things by imagining the relevant situations and giving a judgement on them. Even limiting ourselves to more interesting cases, Gettier cases would be counted by most as clear cases in which the concept KNOWLEDGE fails to apply, despite some of the features of knowledge being present (in that respect, they are similar to cases which are known to create problems for prototype theories – of concepts, or even of their application: a robot-cat might have most of the prototypical features of cats, and yet it is a clear case of non-cat).

Let me now get to the second reply. Obviously, some philosophical cases are hard, as witnessed by the amount of discussion they spawn. Plausibly, they often are platypus cases. The trolley problem could be used as an example here; there is an action which results in the death of an innocent person, and yet it results in avoidance of the death of a number of other innocent people. The former fact is usually sufficient reason to deem the action morally impermissible, and the latter usually sufficient to count the action as permitted and even required. One might even think that more often than not, thought experiments are interesting *because* they are platypus cases, and so they would not be interesting if they were not hard.

What should we conclude from the fact that some thought experiments in philosophy are hard? One might be tempted here to question two assumptions that I have made at the beginning of this chapter, i.e. that there is a determinate answer about certain questions as to whether a certain concept applies to a certain imaginary situation and that disagreement is not

by itself evidence of verbal disagreement. Indeed, this is at least on a reading, what Szabó Gendler is doing. She is arguing that the sort of intractable disagreement that we observe in some cases is good evidence that the concept of person we normally employ does not deliver any determinate result in some of the cases discussed in the literature. Of course, we may create new, more precise concepts, that resemble out ordinary concept and also give a verdict on the platypus cases. But there is no single, context-independent, best way of creating such a concept. If this is the criticism, then I am neutral on its efficacy in specific cases. To evaluate the idea, one would have first to be clear about the logic of a concept which is indeterminate in this way, and then to look at the semantics and metaphysics of the subject-matter under discussion, e.g. persons, or moral norms, and discuss the proposal there. I am not arguing that this kind of indeterminacy could not occur, and I am not even assuming that it does not. I am assuming however that it does not occur so often to make the practice of appealing to judgements about thought experiments defective. There is perhaps a reading of the challenge of experimental philosophy on which the experimental philosophers are trying to give empirical support to this kind of 'indeterminism'. I take it that such a project to be very different from, and arguably incompatible with[103], the sort of challenge I discuss here, which is also the most prominent in the experimental philosophy literature, and is a challenge to the reliability on one's judgments about hypothetical cases. Note again that the main motivation for the claim that most thought experiments do not have determinate answers would lie not in the empirical evidence, but rather in some general views about meaning. While I do not have space to discuss those views properly in this context, I believe that those views of meaning are misguided, and they are probably related to the confusion between concepts and the ability to apply them I discussed in chapter 4. If we identify conceptions and concepts, we end up with the result that whenever we do not know what to say about a case, because our conception does not give a clear verdict, the concept itself is indeterminate, and its application result in something which is neither true not false. But we have good reasons to think, as I argued in chapter 2, that our concepts have application condition which may radically transcend the capacity of the individuals employing them. If we take care to distinguish concepts and conceptions, the fact that a case puts pressure on our conception is at best weak evidence that the concept itself is indeterminate. From now on, I am going to only consider

[103] It is at best unclear whether questions about reliability make sense in absence of the possibility of true, or false, judgments.

the version of the challenge that grants the existence of a correct judgement about thought experiments, but questions our reliability in reaching it.

So far we have concluded that there are hard cases, but, I am assuming, at least some of them still have a determinate answer. Nothing was found that justifies a worry about philosophy in particular. Hard cases are hard, and there are some in every discipline. However, one might claim, the point is precisely that, while as a matter of fact philosophers often use their first judgment about complicated thought experiments as evidence, when a philosopher is confronted with such a case, she should *not* use her judgment on the case as evidence, since her judgment will not be reliable. But nothing we said so far motivates the latter claim; it does not follow from the fact that a case is hard that philosophers cannot be often right about the judgements to be given about it, and so it does not follow that the practice of appealing to thought experiments as it is now needs any change. The two latter conclusions would only follow if we had reasons to believe some additional empirical facts about philosophers and their practice. One is that philosophers do not possess the kind of expertise that at least partially compensates for the hardness of the cases. A second assumption is that philosophers do not engage in additional checking before accepting a judgement on a difficult case. I think there are some reasons to think these claims are false. As to the second assumption, I would say it is obviously false. Focusing on Gettier cases might be in *this* respect misleading; one exceptional feature of that (kind of) case is that it had a deep theoretical impact in a quick and, as it were, revolutionary way. Most interesting thought experiments are not treated like that. Theorists whose views are threatened by the thought experiment usually try to dismiss initial judgements on the case: arguments are offered, and long discussions take place. As to the second assumption, the idea that philosophy provides any sort of expertise in dealing with thought experiments is the object of an ongoing discussion. Clearly, it is an entirely empirical matter whether this is true. I will deal more thoroughly with this issue in the next section.

If in fact philosophers did not possess any sort of expertise and they proceeded using unreflective judgements on hard cases as evidence, this would indeed be a problem, although it would not involve all thought experiments in philosophy, and not even all interesting ones. Note how the problem, even making such pessimistic and (I am going to argue) unjustified assumptions, would still not affect the main thesis I am defending, which is that thought experiments can provide knowledge, and this only requires an application of our standard ability to employ concepts. So the arguments in the next section, to the extent that they work, provide evidence in favour of a stronger thesis, the thesis that philosophers can acquire

justified beliefs through the use of thought experiments even when the thought experiments present then with hard and interesting cases.

5.3. Expertise

Weinberg et al. [2010] examine the psychological literature on expertise, and argue for the conclusion that "philosophers have no reason to expect now, and from the armchair, that we are intuitive experts of the right sort" (p. 350). I am not sure what one should expect "from the armchair"; but I think that given what we know about the way philosophy is practiced (here I am thinking mainly of philosophy in the analytic tradition), and given what we know from the studies in psychology about expertise, we do have some reason for optimism, at least enough to allow a philosopher to make use of thought experiments and reach justified conclusions. Let me stress however one way in which I do not disagree with Weinberg et al.; I do not argue for the claim that more empirical research on the matter is utterly unjustified. In fact, I even suggest a method, taken from the literature in psychology, which is best suited for conducting such research. However, the fact that more research on the reliability of a method is not a waste of time does not entail that there is anything wrong with that method. It might be worth reminding the reader that in general we do not need, in order to reach justified beliefs using a certain method, to have established scientifically that the method is reliable; this requirement would start an infinite regress.

5.3.1. Two Forms of Philosophical Expertise

Weinberg et al. point to three kinds of factors which might explain what constitutes the expertise in evaluating thought experiments which one acquires through philosophical training. They list these factors as follows: "The three hypotheses that we will consider are that philosophers have superior *conceptual schemata* to the folk; that they deploy more sophisticated *theories* than the folk; and that they possess a more finely-tuned set of *cognitive skills* than the folk." (p. 336, orig. italics). They then argue that there is no reason to think that any of these factors is actually at work, either because there is no good explanation of how it would help in this case, or there is no explanation of how philosophers could acquire the relevant trait, or it is question-begging to assume that they do (or some combination of these). I will concede, for the sake of the argument, that supposing philosophers deploy more accurate explicit theories of the subject matter will not be of help here; mostly because it is not clear we apply any *explicit* theory, at least in many cases of judgments on a thought experiment. I think we don't need to worry about this, because the other two points clearly

provide us with a rich account of how philosophers are likely to apply the relevant expertise. Let me consider them in turn.

5.3.1.1.Conceptual Schemata

Weinberg et al. refer to Camerer and Johnson [1991] as the main source of the idea that experts in different fields employ a sort of 'conceptual schemata', which is developed through practice, and differentiates them from novices. This could be realized in a 'configural rule', a sort of localized heuristic method to form judgment on specific kinds of cases. This hypothesis conforms extremely well to the theory advanced in chapter 4. The configural rules would be part of tacit theory or a prototype, which I think of as a capacity to employ a concept.

It is important to note the crucial role of specialization on this model – experts in one sub-field of analytic philosophy will, in this respect, have a distinct advantage at dealing with imaginary cases that use concepts from their area of specialization. Take the case of epistemology; practicing epistemologists will have years of experience in reflecting on the distinctions between, e.g., justification, knowledge, belief, truth, evidence, rationality, reliability, and so on. These distinctions are developed through reflection on other imaginary cases, but not only that. Explicit definition, formal tools (epistemic logic, probability theory), reflection on actual cases (both from history of science and everyday experience) seem to be equally important. Philosophers who specialize in different areas might not have developed the same ability to apply this specific concepts.

The main objection Weinberg et al. [2010] raise against this hypothesis is that it is unclear how the philosophers could develop this ability in a reliable way, in the absence of reliable feedback on their previous applications of the concepts.

In part, I have already suggested an answer, by pointing to the fact that philosophers not only reflect on imaginary cases. But it is also worth noting that the target problem here is relevant again. Not all imaginary cases are terribly hard; Weinberg et al. [2001] for example use a clear case of absence of knowledge as a sort of control, and they found that their subjects answered correctly on that case, independently of cultural or socio-economical factors. That is one way in which philosophers get reliable feedback. Any introduction to a philosophical sub-field is likely to contain a variety of such easy cases. But, one might object, how could being exposed to easy cases help one to build the ability to deal with hard ones? The problem with hard cases, the objection would go, is precisely that they are unlike the easy ones, and often our usual procedures to decide whether a concept applies, both implicit and

explicit, will be useless in a far-fetched thought experiment. There is of course something right in this thought. But note first that the problem is not more pressing here than in other fields. If we need to decide what the correct explanation of a patient's symptoms is, or what the correct move is in certain chess situation, and so on, we will still have hard cases and easy ones; and the training of a novice will usually involve many cases of the latter kind. Still, one might ask, *how* is it that experience with easy cases helps one in acquiring the kind of expertise helps with the hard ones? The answer, according to the model of expertise we are considering here, is that dealing with a wide range of different cases (not all easy cases are alike) helps in developing richer conceptual schemata, and the richer conceptual schemata may help where a simpler one would be stuck. We can illustrate this point again through the paradigmatic example of the Gettier cases. It is possible that reflection on different cases of true belief that clearly fall short of knowledge, and others that clearly constitute knowledge, brings to the development of some configural rules which in turn deliver the result that the subject in a Gettier case does not know. Plausible candidates are a "no false lemmas" rule, or a reliability rule. According to the former, a true belief is not knowledge if it is inferred from a false belief; according to the latter, a belief is not knowledge if it is produced by an unreliable process. These are both plausible generalizations (although at least the former might have exceptions, as noted in the previous chapter) which can be acquired through familiarity with cases of unjustified true beliefs, and then applied to the unfamiliar sort of justified true belief we encounter in Gettier cases. Such rules of course have been later made explicit and proposed as explanations of the lack of knowledge of the subject in a Gettier case; the hypothesis that the rules were implicitly used in reaching a judgment is of course independent of the theoretical validity of such proposals. The only claim I am defending here is that those rules could provide the correct verdict about some Gettier cases.

5.3.1.2. Know-How with Hypothetical Scenarios

Reading the text of a thought experiment and understanding it requires a certain amount of cognitive skill. A related capacity which might be relevant is distinguishing pragmatic and semantic factors in the appropriateness of a statement. This is one important skill that, according to Weinberg et al., philosophers possess; but they fail to see its relevance (they consider it under the 'conceptual schemata' heading, while it seems to me this capacity is part of general know-how with interpreting the description of a scenario). They write:

> Use/mention, epistemological/metaphysical, semantic/pragmatic — we certainly think that philosophical training can inculcate expert conceptual schemata structured in terms of these

> dimensions. But the explicitness and clarity of these distinctions, and of our tools for dealing with them, stands in very sharp contrast to the complete inarticulateness of the . . . well, whatever it would be, that is supposed to help trained philosophers to categorize Gettier cases as non-knowledge, and to be insensitive to cultural biases and framing effects. (Weinberg et.al.[2009] p.342)

However, in a footnote (Weinberg et al. [2010] fn. 12) they acknowledge that explanations of the data in terms of a semantic/pragmatic distinction were advanced both about Machery et al. [2004] and about part of the results in Weinberg et al. [2001]; they also acknowledge that in the latter case the explanation seems very likely to be correct. Given that the two papers just mentioned are often considered as paradigmatic examples of what experimental philosophy is, and in particular of the way experimental philosophy poses a challenge to the traditional 'armchair' methods of philosophy (see for example their prominent collocation in Knobe and Nichols [2008]), it seems at least part of "the ... whatever it would be, that is supposed to help trained philosophers" is just staring the authors in the face, but they refuse to look at it.

Even the mere fact of being used to consider imaginary cases and reflecting about them should be expected to help, other things being equal. Sorensen [1992], well before the experimentalist challenge was brought to the attention of the philosophical community, argues that the use of bizarre or far-fetched scenarios to refute definitions and necessity claims is an 'anti-fallacy'; an anti-fallacy is an inference pattern which is valid, but is often mistakenly rejected as invalid. He considers various interpretations of the 'bizarreness' claim, and concludes that none really presents us with a distinct epistemological problem (Sorensen [1992] pp. 274-283). It is worth noting that one way in which philosophers might avoid this 'anti-fallacy' is by distinguishing sharply different sorts of possibility; what is epistemically or practically impossible might be nomologically or metaphysically possible, and so on. A different way in which philosophers might be better off is in being clearer about their target; a bizarre or far-fetched case does not refute a generalization or a *ceteris paribus* causal claim, but it does refute a definition or a necessity claim. Correspondingly, most often the significant cognitive achievement in a philosophical counterexample is not giving the correct judgment about the hypothetical case, but rather thinking about the case and noting its relevance for the debate at hand; the Gettier case is paradigmatic in this sense. Gettier's achievement, in short, was not seeing that in his case the subject does not know, but rather coming up with the case and using it as a counterexample to the justified true belief theory.

As for the practice of dealing with difficult cases, philosophers in the interesting cases are used to reflect about their initial judgements, comparing them with those given by other philosophers, checking whether they are in contrast with other justified beliefs that they have,

and whether there are reasons to abandon the initial judgement in favour of the previous beliefs, particularly when arguments are offered to that effect; moreover, it is standard to consider various similar cases and assessing what factors are influencing the relevant judgements. In particular it seems philosophers are usually aware of the possibility that framing effects are at work, and whenever the judgment involved is controversial they try to re-describe the case to check for its stability [104].

5.3.2. Expertise in the Real World

In this section, I will argue that Weinberg et al. [2010] fail to consider some parts of the psychological literature on expertise which are not favourable to their thesis.

An idea which I will not insist on is that there is a correlation between the time spent in training and the acquisition of expertise. In particular, sometimes this correlation has been given expression by the so-called "ten-years rule" (see Ericsson [1996]; Simon and Chase [1973] first introduced the idea). The ten-years rule, which was thought to be surprisingly stable across different domains, states that acquiring expertise in an area requires roughly ten years, or ten thousands hours, of deliberate practice. However, more recent research suggests that this is at best a necessary condition on the acquisition of expertise.

There is however a more interesting and up-to-date approach to the evaluation of expertise which should be considered, and that is the Cochran-Weiss-Shanteau (CWS) method (it is peculiar that Weinberg [2009] and Weinberg et al. [2010] make ample reference to Shanteau's older work, and none to his works published after 1992). The CWS method is a method developed with the aim of evaluating expertise in areas in which there is no external standard to evaluate the putative expert performance against (we will come back to this). According to this recently developed method the relevant factors in evaluating a putative expert are just two: consistency and discrimination. Here is a first explication of these two features: "Discrimination refers to a judge's differential evaluation of different stimulus cases. Consistency refers to a judge's evaluation of the same stimulus over time; inconsistency is its complement." (Shanteau et al. [2002], p. 257)

According to the view, each of the two factors is necessary, and together they represent a positive indicator of expertise (although maybe not a sufficient condition in a

[104] A paradigmatic example of that is Bernard Williams' re-description of a case which was supposed to support the thesis that bodily continuity is not essential to personal identity, to get the opposite 'intuitive' judgment (in Williams [1970]. One might also think of Thomson variations on the original Trolley cases as instances of framing effects. I am indebted here to Cappelen [forthcoming].

strong sense). The authors propose that we obtain a good indication of the degree of expertise by the ratio between discrimination and inconsistency, which we may call CWS ratio. The CWS ratio is obtained by dividing a value obtained for discrimination by a value obtained for inconsistency. The idea behind the CWS method is explained as follows:

> The intuition underlying the index is that a good measuring tool necessarily has a high CWS ratio. That is, a proper instrument yields different measures for different objects, and gives the same measure whenever it is applied to the same object. A ruler, for example, discriminates among objects of varying length, and produces identical scores for the same objects. (...)
>
> Similarly, an expert must be both discriminating and consistent. It is easy to display one or the other, but hard to do both. One can show discrimination by generating wide variety of responses over stimuli; one can exhibit consistency by repeating the same response to all stimuli. But adopting either of these strategies alone means that the other entity will be lost. To display both properties simultaneously requires careful assessment of the stimuli, the essence of expert judgment. (Shanteau et al. [2002] p. 258)

The appropriateness of the method however need not rest on these considerations alone. The method can be tested empirically (Shanteau et al. [2002] is a survey of previous studies, Weiss et al. [2009] describes some recently conducted targeted studies; see further references there). The CWS method has been evaluated applying it to (putative) experts' judgments or performances for which we have an independent measure of success (including medical diagnostics, weather forecast, mental calculations, simulations of air traffic control and many others). The results so far have been very encouraging.

This is not the place for an accurate assessment of the empirical adequacy of the CWS method. More studies are needed, as the proponents of the method acknowledge. However, it seems that the arguments and the evidence in favour of the CWS method are strong enough to make it interesting to consider in relation to philosophy. After all, the whole psychological literature on expertise is a fairly young research filed, and very few significant results are fully established. Moreover, the CWS method is particularly well-suited for our purposes. As mentioned above, the method was developed with the aim of evaluating expertise in areas in which there is no external standard to evaluate the putative expert performance against, based only on the observation of the expert, as it were. Strictly speaking, this is not completely true, since we need a criterion to decide whether the case judged is or is not the same. Still, this is certainly much easier than telling what the correct judgment about a case is. It is only required that we individuate sufficient conditions for two cases to represent a constant stimulus. With respect to philosophical thought experiments, we can normally assume that a text describing a scenario represents a constant stimulus. Failure to give the same judgement on identical texts

will be a failure of consistency. This is a simplification, because of problems related to context-sensitivity; where context-sensitive terms are involved, the same text can actually describe different scenarios. This will be problematic when what is at stake is precisely the context-sensitivity of some term involved (as in debates about the context-sensitivity of "knowledge"), and it can be an explanation for unexpected order effects even when there is no initial presumption that context-sensitivity is involved[105]. Nonetheless, it should be clearly possible to judge of the sameness of the scenarios involved in a vast number of cases, independently of what the correct judgements on the scenario are. Moreover, there are several cases in which certain variations in the description of a case (e.g. the colour of the dress someone wears, or the gender of someone involved in a scenario) are recognized as irrelevant without controversy; so cases that only differ in those aspects can be counted as the same.

The CWS method is thus particularly well-suited for those who wish to assess empirically the possession by philosophers of expertise in making judgments about thought experiments. Barring the problems noted at the end of section 5.1, it seems we have here a useful suggestion for experimental philosophers interested in this particular problem. However, at present I can only give a judgment from the armchair, as it were (not a priori, of course, on any reading of the expression, because the judgment is informed by my past experience with philosophers and non-philosophers), about how the philosophers would score on a CWS index. It seems not at all unlikely that philosophers will have a far better CWS index then non-philosophers. In particular, it seems likely that they will have a higher degree of consistency, and a similar degree of discrimination. While I do not wish to put much weight on this armchair empirical prediction, even the conditional claim that if the prediction is correct, then we have a good indication of expertise, is important to keep in mind, since Weinberg et al. [2010] dispute it. They write

> Although we've no evidence for it, we will also grant, for the sake of the argument, that the philosophical intuitions of experts will stabilize on some set of verdicts in such a way that they are able to give the same answer across conditions in the standard restrictionist experiments. We'll grant this because we don't think it makes much difference to the dialectic. (Weinberg et al. [2010], p. 339)

The reason they do not think this would make much difference is that one would need, they think, an independent argument to establish that such consistency is achieved through a

[105] Knobe and Szabó [manuscript] present a very interesting hypothesis which explains as semantic context-dependence certain patterns of judgements which were previously attributed to order effects or influence of irrelevant factors.

reliable method, and not through some arbitrary rule, for example following a theory which is itself based on an initial set of unreliable intuitions. This is a dangerous step for Weinberg and colleagues, for they risk to say that actually any empirical study is irrelevant here; they will just think that philosophers judgments on thought experiments are unreliable even they do not show any dependence on the irrelevant factors that folk judgments depend on. This is a terrible step from the dialectical point of view, for at the same time they beg the question and they run against the spirit of experimental philosophy. Luckily, they have a good reason to avoid this step. If philosophers show better consistency, provided they score at least as good as non-philosophers on discrimination, then there are good reasons to think that they have genuine expertise.

There are two more interesting lessons to be learned from the literature on the CWS method, quite independently of how philosophers will perform on this standard. The first has to do with the observation just made that the method is aimed at evaluating expertise in areas in which no external standard is available. Weinberg et al. [2010] sometimes write as if it was a peculiar and fatal disadvantage of philosophy that there are no ways of assessing judgments about hypothetical cases, of telling which of them are correct, except those that rely directly or indirectly on the philosophers' (expert) judgment. For example:

> Another suggestion might be that philosophers train their intuitions against other, already-certified expert intuitions. But this appears to be a non-starter, since it just invites an explanatory regress: how did the purveyors of those intuitions develop their expertise? (p. 341)

The CWS method is aimed precisely at measuring expertise in fields in which there is no external standard on which to assess the (alleged) expert performance. Defenders of this method convincingly argue at several points that this is the typical situation in which we are with respect to evaluating experts. I think it is useful to give some quotes.

> Weiss and Shanteau [2004]: The reason we need experts in the first place is that they offer us answers that we could not obtain any other way (p.231)

> Shanteau et al. [2002]: Indeed, if we could compute (or look up) a correct answer, why would we need an expert at all? (p.253)

> Weiss et al. [2009]: What is the outcome that reflects the quality of a film review, the grade assigned by an instructor, or the sentence imposed by a magistrate? (p.165)

> Weiss and Shanteau [2003]: For many tasks at which experts make a living, no measurable outcome exists. How is one to know if the wine taster has judged accurately or if the professor has graded the essays well? (…) Although there is no hint of an objective external criterion, we

> believe that some people do these tasks better than others and that people improve their performance. We would like our assessment scheme to include such expertise. (p.105)

If it were impossible to develop expertise based without having a feedback other than other experts' judgments, we would have to give up on the idea that there is expertise in any of the areas mentioned in these quotes; in other words, we would have to give up on the idea that there is expertise, except where expertise is not needed.

The second lesson has to do with the irrelevance of agreement among (putative) experts. Another common piece of the rhetoric against the use of thought experiments in philosophy (or against philosophy as a whole) has to do with the lack of consensus among philosophers on the correct judgments. Defenders of the CWS method stress how consensus, i.e. being in accordance with other (socially recognised) experts, is not even a necessary condition on someone being an expert (see especially Shanteau [2000] and Weiss and Shanteau [2004]). Of course, where experts differ in their judgments, some of them are wrong (there is no reason here not to employ classical logic). For example, if the experts are split in two equally numerous groups on a yes/no question, then half of them are going to be wrong. Now, one is tempted to reason, this means that experts are only 50% reliable, so even the experts who are right are not reliable enough to be trusted. But this reasoning is fallacious. There is a scope ambiguity, we might say, in "experts are 50% reliable". If it means that picking an expert at random you have 50% chance of that expert giving a right answer, then (in the case imagined) that is correct. But if it means that each expert has a 50% chance of giving a correct answer, then clearly it says something that is not implied by the situation described. It could be that the experts who have the correct answer reached it through a completely reliable process. In such a situation, the CWS method promises to give us a way to evaluate which group is more reliable. Moreover, Weiss and Shanteau [2004] argue that this can be the case even for experts defending a minority view. To make this vivid, think of a moment in the history of science in which geocentrism was the majority view, but a minority of Copernicans already existed. It is surely possible that the minority was using a reliable method. Consensus in a scientific community, given that the judgments are not reached independently, is not a sure sign of reliability, and conversely, even a view defended by a single expert may have been reached through a reliable process. Of course, in cases in which the correct view is held by a minority of socially recognized experts, it will be hard to judge from the point of view of non-experts that this is the case. However, there is a lesson here too;

when experts disagree, simply counting is not a good way of deciding which experts we should trust[106].

Overall, it seems that close attention to this part of the psychological literature on expertise makes the arguments in Weinberg et al. [2010] appear mostly as unjustified armchair scepticism about philosophy, supported by an aggressive but ultimately unconvincing rhetoric.

Conclusion

I have considered an argument, presented explicitly in Weinberg et al. [2010], and implicitly in the previous literature in experimental philosophy, to the effect that philosophers are not reliable in their judgments about a certain class of thought experiments. The argument proceeds from an empirical premise about the unreliability of non-philosophers, and extends that result, inductively, to philosophers. In the second part of the chapter, I have argued that there is no way of individuating the class of thought experiments under discussion which makes the argument both interesting and sound. In the last section, I have argued, on the basis of both the discussion in Weinberg et al. [2010] and other parts of the psychological literature on expertise, that the inductive step in the argument has not been justified; there are reasons to think that philosophers possess some sort of expertise in evaluating thought experiments.

[106] Goldman [2001] reaches a similar conclusion, and also advances several hypotheses as to what other factors are relevant from the point of view of non-experts to assess the reliability of experts.

Conclusion

Here is a brief summary of what I believe has been achieved in this work, and some remarks about what I believe has not been achieved, but might be in the future.

In the first three chapters I discussed and rejected several views about philosophical methodology which are rather popular. In order of increasing popularity, I argued against the view that intuitions, as a *sui generis* mental state, are involved crucially in philosophical methodology (chapter 1); the view that philosophy requires engagement in conceptual analysis, understood as the activity of considering thought experiments with the aim of throwing light on the nature of our concepts (chapter 2); and the view that much philosophical knowledge is a priori (chapter 3). I believe that I have shown, at the very least, that these views are much more problematic than many people realize, and that there is fundamental work to be done for their defenders, if they are to be rescued. Of course, I am sceptical about the prospects of success of these projects. I have considered several versions, usually prominent ones, of each of the views, and I have shown those versions to be defective. Quite often, different versions of the same worry applied to different versions of the same theory. I do not claim to have a proof that nothing in the vicinity of these views is correct; such a proof might well be impossible to give, unless some constraint can be given on how we need to evaluate, not on a case by case basis, what counts as a relevantly similar view; and I do not see how what form such constraint could take.

In the last two chapters I developed a more positive account of (part of) philosophical methodology. I defended Williamson's view that the crucial step in our reasoning about a hypothetical case is best represented as a counterfactual judgement. I then discussed the epistemology of the judgements involved in philosophical thought experiments, arguing that their justification depends on their being the product of a competence in applying the concepts involved, a competence which goes beyond the possession of the concepts. I then offered, drawing from empirical psychology, a sketch of the form this cognitive competence could take. I argued that the overall picture is plausible and it squares well with the conclusions of the first part. In the last chapter I considered an argument against the use of thought

experiments in contemporary analytic philosophy found in authors who belong to the 'experimental philosophy' movement. The argument proceeds from an empirical premise about the unreliability of non-philosophers about a certain class of cases and extends that result, by analogical reasoning, to philosophers. I argued that there is no way of individuating the class of thought experiments under discussion which makes the argument both interesting and sound. Moreover, I argued that the analogical step in the argument has not been justified; there are reasons to think that philosophers possess several kinds of expertise which set them apart from non-philosophers in relevant ways.

There are at least two directions in which the discussion, I believe, should be (and perhaps, ideally, should have been) extended. These are the investigation of modal epistemology and of the epistemology of logic and mathematics. Of course, modal reasoning and logical and mathematical reasoning are ubiquitous outside philosophy as well; they constitute a large part of any kind of investigation. However, clearly they are topics that many philosophers have been often concerned with, hence their epistemology overlaps with the epistemology of philosophy. I believe many of the things I wrote have obvious bearing on the epistemology of these subjects. In particular, I believe that the negative arguments in the first three chapters (perhaps more clearly for the first two chapters than for the third) are, by and large, sufficiently general to apply to those areas as well. Of course this would need to be shown in more detail. On the other hand, the positive account of knowledge about hypothetical cases offered in chapter 4 would have to be developed substantially to be applied in these areas. The main difficulty I see in applying a theory of conceptions to these topics is that the conceptions themselves plausibly contain modal and logical elements. It is part of our implicit theory of at least some things, or kinds, that certain attributes are more essential to them than others. Moreover, in order for a conception to be applied in any specific case some kind of implicit inference has to be drawn, and therefore some logical competence must regulate it. The upshot is that appealing to conceptions by itself might not be of great help in understanding the epistemology of modality or logic, since conceptions incorporate knowledge of the sort that we seek to explain.

I would like to end with a final thought about the methodology of philosophical methodology. As I said at the outset, I believe that philosophical methodology is just part of philosophy. Hence, everything I said about the methodology of philosophy should apply, other things being equal, to the methodology of the methodology of philosophy as a particular case. I hope to have been consistent in this sense. My methodology does not seem to me to involve special mental states, analytic truths, or claims justified completely independently of

experience. It did frequently involve appeal to judgements about certain hypothetical cases; none of the cases, as far as I can see, was far-fetched in a way that should cause deep worries. No doubt, those who disagree with my conclusions about philosophical methodology may also disagree about the description I just gave of my own methodology. To revert to Neurath's metaphor, made famous by Quine, those of us who engage in philosophical methodology are like sailors who must rebuild their boat while in the open sea, without any safer place to stand on while we are working than the boat itself. This is perhaps not an ideal situation, but it is certainly better than refusing to look closely at how the boat is working, as some radical rationalists seem to do, or trying to repair its allegedly serious leaks by sinking it altogether, and us with it, as some radical experimentalists seem to do.

Bibliography

Armstrong, Sharon L., Gleitman, Lila R. and Gleitman, Henry [1983], "What Some Concepts Might Not Be", *Cognition* 13, pp. 263-308.

Audi, Robert [1989], "Causalist Internalism", *American Philosophical Quarterly* 26, pp. 309-20.

Barsalou, Lawrence W. [1987], "The Instability of Graded Structure: Implications for the Nature of Concepts", in U. Neisser (ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization,* Cambridge University Press, pp. 101-40.

Bealer, George [1992], "The Incoherence of Empiricism", *Proceedings of the Aristotelian Society* 66, pp. 99-138.

- [1996], "On the Possibility of Philosophical Knowledge", *Philosophical Perspectives* 10, *Metaphysics,* pp. 1-34.

- [1998], "Intuition and the Autonomy of Philosophy", in M. DePaul and W. Ramsey (eds.), *Rethinking Intuition,* Rowman & Littlefield Publishers Inc., pp. 201-40.

- [2000], "A Theory of the A Priori," *Pacific Philosophical Quarterly* 81, pp. 1-30.

- [2002], "Modal Epistemology and the Rationalist Renaissance", in T. Szabó Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*, Oxford University Press, pp. 71-125.

- [2004], "The Origins of Modal Error", *Dialectica* 58, pp. 11-41.

Block, Ned and Stalnaker, Robert [1999], "Conceptual Analysis, Dualism, and the Explanatory Gap", *Philosophical Review* 108, pp. 1-46.

Boghossian, Paul [1989], "Content and Self-Knowledge", *Philosophical Topics* 17, pp. 5-26.

- [1996], "Analyticity Reconsidered", *Nous* 30, pp. 360-91.

- [1998], "What the Externalist Can Know A Priori", in C. Wright, B. Smith and C. MacDonalds (eds.), *Knowing Our Own Minds*, Clarendon Press, pp. 271-84.

- [2003a], "Blind Reasoning", *Aristotelian Society Supplementary Volume* 77, pp. 225-48.

- [2003b], "Epistemic Analyticity: A Defense", *Grazer Philosophische Studien* 66, pp. 15-35.

- [2009], "Virtuous Intuitions: Comments on Lecture 3 of Ernest Sosa's *A Virtue Epistemology*", *Philosophical Studies* 144, pp. 111-19.

- [2011a], "Review of *Truth in Virtue of Meaning*, by Gillian Russell, *Australasian Journal of Philosophy* 89, 370-374

- [2011b], "Williamson on the *A priori* and the Analytic", *Philosophy and Phenomenological Research* 82, pp. 488-97.

BonJour, Laurence [1998], *In Defense of Pure Reason: a Rationalist Account of A Priori Justification*, Cambridge University Press.

- [2001], "Reply to Harman", *Philosophy and Phenomenological Research* 63, pp. 691-5.

- [2005], "In Defense of the *A Priori*" in M. Steup and E. Sosa (eds.), *Contemporary Debates in Epistemology*, Blackwell Publishing.

Brown, James Robert [2007], "Thought Experiments", *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/thought-experiment/

Brown, Jessica [1995], 'The Incompatibility of Anti-Individualism and Privileged Access', *Analysis* 55, pp. 149-56.

- [2001], "Anti-Individualism and Agnosticism", *Analysis* 61, pp. 213-24.

- [2004], *Anti-Individualism and Knowledge*, MIT Press.

Brueckner, Anthony [1992], "What an Anti-Individualist Knows A priori", *Analysis* 52, pp. 111-18.

Burge, Tyler [1979], "Individualism and the Mental", *Midwest Studies in Philosophy* 4, pp. 73-122.

- [1982a], "Other Bodies", in A. Woodfield (ed.), *Thought and Object*, Oxford University Press.

- [1982b], "Two Thought Experiments Reviewed", *Notre Dame Journal of Formal Logic* 23, pp. 284-93.

- [1986], "Intellectual Norms and Foundations of Mind", *Journal of Philosophy* 83, pp. 697-720.

- [1990], "Frege on Sense and Linguistic Meaning" in D. Bell and N. Cooper (eds.), *The Analytic Tradition*, Blackwell.

- [1993a], "Content Preservation", *Philosophical Review* 102, pp. 457-88.

- [1993b], "Concepts, Definitions and Meaning", *Metaphilosophy* 24, pp. 309-25.

- [1997], "Interlocution, Perception and Memory", *Philosophical Studies* 86, pp. 21-47.

- [1998], "Computer Proof, A Priori Knowledge, and Other Minds", *Philosophical Perspectives* 12, pp. 1-37.

- [2010], *Origins of Objectivity*, Oxford University Press.

Cappelen, Herman [forthcoming], *Philosophy without Intuitions*, forthcoming for Oxford University Press

Carey, Susan [1991], "Knowledge Acquisition: Enrichment or Conceptual Change?", in S. Carey and R. Gelman (eds.), *The Epigenesis of Mind: Essays on Biology and Cognition*, Lawrence Erlbaum Associates, Inc. Publishers, pp. 257-91. Reprinted in Margolis and Laurence [1999a].

- [2009], *The Origin of Concepts*, Oxford University Press.

Casullo, Albert [2003], *A Priori Justification*, Oxford University Press.

Chalmers, David [2002], "The Components of Content", in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, Oxford University Press, pp. 608-33.

- [2006a], "The Foundations of Two-dimensional Semantics", in M. Garcia Carpintero (ed.), *Two-Dimensional Semantics*, Oxford University Press, pp. 55-141.

- [2006b], "Two-Dimensional Semantics", in E. Lepore and B. Smith (eds.), *Oxford Handbook of Philosophy of Language*, Oxford University Press, pp. 574-606.

- [forthcoming], "Propositions and Attitude Ascriptions: a Fregean Account", *Nous*.

Chalmers, David and Jackson, Frank [2001], "Conceptual Analysis and Reductive Explanation", *Philosophical Review* 110, pp. 315-361.

Chudnoff, Elijah [2011], "The Nature of Intuitive Justification", *Philosophical Studies* 153, pp. 313-33.

Coffman, E. J. [2008], "Warrant without Truth?", *Synthese* 162, pp. 173-94.

Cummins, Robert [1998], "Reflections on Reflective Equilibrium", in M. DePaul and W. Ramsey (eds.), *Rethinking Intuition*, Rowman & Littlefield Publishers Inc., pp. 113-27.

Dancy, Jonathan [1985], *An Introduction to Contemporary Epistemology*, Basil Blackwell.

DePaul, Michael [2009], "Phenomenal Conservatism and Self-Defeat", *Philosophy and Phenomenological Research* 78, pp. 205-212.

DePaul, Michael R. and Ramsey, William (eds.), [1998], *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*, Rowman and Littlefield Publishers Inc.

Devitt, Michael [2005], "There is no *A Priori*", in M. Steup and E. Sosa (eds.), *Contemporary Debates in Epistemology*, Blackwell Publishing, pp. 105-15.

Ericsson, K. Anders [1996], "The Acquisition of Expert Performance: An Introduction to Some of the Issues", in K. A. Ericsson (ed.), *The Road to Expert Performance: Empirical Evidence from the Arts and Sciences, Sports and Games*, Erlbaum, pp. 1-50.

- [2006], "The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance", in K. A. Ericsson, N. Charness, P.J. Feltovich, R.R. Hoffman (eds.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press.

Evans, Gareth [1979], "Reference and Contingency", *The Monist* 62, pp. 161-189.

Evans, Jonathan St. B. T. [2007], *Hypothetical Reasoning. Dual Processes in Reasoning and Judgement*, Psychology Press.

Fara, Michael [2005], "Dispositions and Habituals", *Nous* 39, pp. 43-82.

- [2006], "Dispositions", *Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/entries/dispositions/

Field, Hartry [2005], "Recent Debates about the A Priori", in T. Szabó Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology, vol. 1*, Oxford University Press, pp. 69-88.

Firth, Roderick [1978], "Are Epistemic Concepts Reducible to Ethical Concepts?", in A. I. Goldman and J. Kim (eds.), *Values and Morals*, Reidel, pp. 215-29.

Fodor, Jerry [1998], *Concepts: Where Cognitive Science Went Wrong*, Clarendon Press.

- [2008], *LOT 2: The Language of Thought Revisited*, Oxford University Press.

Frankish, Keith and Evans, Jonathan [2009], "The Duality of Mind: An Historical Perspective", in J. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond*, Oxford University Press, pp. 1-29.

Gelman, Susan A. and Wellman, Henry M. [1991], "Insides and Essences: Early Understanding of the Non-Obvious", *Cognition* 38, pp. 213-44. Reprinted in Margolis and Laurence [1999a]

Goldberg, Sanford [2003a], "Anti-Individualism, Conceptual Omniscience, and Skepticism," *Philosophical Studies* 116, pp. 53–78.

- [2003b], "On Our Alleged *A Priori* Knowledge That Water Exists," *Analysis* 63, pp. 38–41.

Goldman, Alvin [1986], *Epistemology and Cognition*, Harvard University Press.

- [1999], "A Priori Knowledge and Naturalistic Epistemology", *Philosophical Perspectives* 13, *Epistemology*, pp. 1-28.

- [2001], "Experts: Which Ones Should You Trust?", *Philosophy and Phenomenological Research* 63, pp. 85-110.

- [2002], “A Priori Warrant and Naturalistic Epistemology”, in A. Goldman, *Pathways to Knowledge. Private and public*, Oxford University Press, pp. 24-50.

- [2007], “Philosophical Intuitions: their Target, their Source and their Epistemic Status”, *Grazer Philosophische Studien* 74, pp. 1–26.

Goldman, Alvin and Pust, Joel [1998], “Philosophical Theory and Intuitional Evidence”, in M. DePaul and W. Ramsey (eds.), *Rethinking Intuition*, Rowman & Littlefield Publishers Inc., pp. 179-200.

Harman, Gilbert [1973], *Thought*, Princeton University Press.

- [1986], *Change in View: Principles of Reasoning*, MIT Press.

- [2001], “General Foundations versus Rational Insight”, *Philosophy and Phenomenological Research* 63, pp. 657-63.

- [2003], “Skepticism and Foundations”, in S. Luper (ed.) *The Skeptics: Contemporary Essays*, Ashgate Press, pp. 1-11.

Hawthorne, John [2002], “Deeply Contingent A Priori Knowledge’, *Philosophy and Phenomenological Research* 65, pp. 247-269.

- [2004], *Knowledge and Lotteries*, Oxford University Press.

- [2007], “A Priority and Externalism”, in S. Goldberg (ed.), *Internalism and Externalism in Semantics and Epistemology*, Oxford University Press, pp. 201-18.

Henderson, David and Horgan, Terry [2001], “The *A Priori* Isn’t All That It Is Cracked Up to Be, But It Is Something”, *Philosophical Topics* 29, pp. 219–50.

Huemer, Michael [2007], “Compassionate Phenomenal Conservatism”, *Philosophy and Phenomenological Research* 74, pp. 30-55.

Hylton, Peter [2007], *Quine*, Routledge.

Ichikawa, Jonathan [2009], “Knowing the Intuition and Knowing the Counterfactual”, *Philosophical Studies* 145, pp. 435-43.

Ichikawa, Jonathan and Jarvis, Benjamin [2009], “Thought-experiment Intuitions and Truth in Fiction”, *Philosophical Studies* 142 , pp. 221-46.

Jackson, Frank [1998], *From Metaphysics to Ethics: A Defense of Conceptual Analysis*, Oxford University Press.

Jenkins, Carrie S. [2008a], “Modal Knowledge, Counterfactual Knowledge, and the Role of Experience”, *The Philosophical Quarterly* 58, pp. 693-701.

- [2008b], “A Priori Knowledge: Debates and Developments”, *Philosophy Compass* 3/3, pp. 436-50.

Kaplan, David [1989], "Demonstratives", in J. Almog, J. Perry and H.Wettstein (eds.), *Themes from Kaplan*, Oxford University Press, pp. 481-563.

Kitcher, Philip [2000], "A Priori Knowledge Revisited" in P. Boghossian and C. Peacocke (eds.), *New Essays on the A Priori*, Oxford University Press, pp. 65-92.

Klein, Peter [1996], "Warrant, proper function, reliabilism, and defeasibility", in J. Kvanvig (ed.), *Warrant in contemporary epistemology*, Rowman & Littlefield.

- [2008], "Useful False Beliefs", in Q. Smith (ed.), *Epistemology: New Essays*, Oxford University Press, pp. 25-62.

Knobe, Joshua and Gendler Szabó, Zoltan [manuscript], "Impure Modals", http://pantheon.yale.edu/~zs47/documents/ImpureModals.pdf

Knobe, Joshua and Nichols, Shaun (eds.), [2008], *Experimental Philosophy*, Oxford University Press.

Korcz, Keith Allen [1997], "Recent Work on the Basing Relation", *The American Philosophical Quarterly* 34, pp. 171-91.

- [2000], "The Causal-Doxastic Theory of the Basing Relation", *Canadian Journal of Philosophy* 30, pp. 525-50.

- [2006], "The Epistemic Basing Relation", *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/basing-epistemic/

Kripke, Saul [1980], *Naming and Necessity*, Harvard University Press.

Laurence, Stephen and Margolis, Eric (eds.), [1999a], *Concepts. Core Readings*, MIT Press.

- [1999b], "Concepts and Cognitive Science", in E. Margolis and S. Laurence (eds.), *Concepts. Core Readings*, MIT Press, pp. 3-81.

- [2003], "Concepts and Conceptual Analysis", *Philosophy and Phenomenological Research* 67 (2), pp. 253-82.

Lewis, David [1983], *Philosophical Papers, vol. 1*, Oxford University Press.

Ludwig, Kirk [2010], "Intuitions and Relativity", *Philosophical Psychology* 23, pp. 427-45.

Machery, Edouard [2009], *Doing Without Concepts*, Oxford University Press.

Machery, Edouard, Mallon, Ron, Nichols, Shaun and Stich, Stephen [2004], "Semantics, Cross-cultural Style", *Cognition* 92, pp. B1-B12.

Maddy, Penelope [2007], *Second Philosophy. A Naturalistic Method*, Oxford University Press.

Malmgren, Anna-Sara [2006], "Is There A Priori Knowledge By Testimony?", *Philosophical Review* 115, pp. 199-241.

- [forthcoming], "Rationalism and the Content of Intuitive Judgments", forthcoming in *Mind*.

Manley, David and Wasserman, Ryan [2008], "On Linking Dispositions and Conditionals", *Mind* 117, pp. 59-84.

McKinsey, Michael [1991], "Anti-individualism and Privileged Access", *Analysis* 51, pp. 9-16.

- [2007], "Externalism and Privileged Access Are Inconsistent", in B. McLaughlin and J. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*, Blackwell Publishing, pp. 53-66.

McLaughlin, Brian and Tye, Michael [1998], "Externalism, Twin-Earth, and Self-Knowledge", in C. Wright, B. Smith and C. MacDonals (eds.), *Knowing Our Own Minds*, Clarendon Press, pp. 285-320.

Nimtz, Christian [2010], "Philosophical Thought Experiments as Exercises in Conceptual Analysis", *Grazer Philosophische Studien* 81, pp. 191–216.

Owens, Joseph [2003], "Anti-individualism, Indexicality, and Character", in M. Hahn and B. Ramberg (eds.), *Reflections and Replies. Essays on the Philosophy of Tyler Burge*, MIT Press, pp. 77-99.

Papineau, David [2009], "The Poverty of Analysis", *Proceedings of the Aristotelian Society Supplementary Volume* 83, pp. 1-30.

Peacocke, Christopher [1992], *A Study of Concepts*, MIT Press.

- [2000], "Explaining the A Priori: the Programme of Moderate Rationalism", in P. Boghossian and C. Peacocke (eds.), *New Essays on the A Priori*, Oxford University Press, pp. 255-85.

Plantinga, Alvin [1993], *Warrant and Proper Function*, Oxford University Press.

Predelli, Stefano [2006], "The Problem with Token-Reflexivity", *Synthese* 148, pp. 5-29.

Pritchard, Duncan [2005], *Epistemic Luck*, Oxford University Press.

Pust, Joel [2000], *Intuitions as Evidence*, Garland Publishing.

Putnam, Hilary [1962], "The Analytic and the Synthetic", reprinted in [1975], *Mind, Language and Reality, Philosophical Papers Vol. 2*, Cambridge University Press.

Quine, Willard Van Orman [1960], *Word and Object*, MIT Press.

- [1970], *Philosophy of Logic*, Prentice Hall.

- [1995], *From Stimulus to Science*, Harvard University Press.

Quine, Willard Van Orman and Ullian, Joseph [1970], *The Web of Belief*, Random House.

Rey, George [1983], "Concepts and Stereotypes", *Cognition* 15, pp. 237-62. Reprinted in E. Margolis and S. Laurence [1999a].

- [2009], "Review of Edouard Machery, *Doing Without Concepts*", *Notre Dame Philosophical Reviews*, online at http://ndpr.nd.edu

Roca-Royes, Sonia [forthcoming], "Modal Knowledge and Counterfactual Knowledge", forthcoming in *Logique et Analyse*.

Rosch, Eleanor [1973], "Natural categories", *Cognitive psychology* 4, pp. 328-50.

- [1978], "Principles of Categorization", in E. Rosch and B.B. Lloyd (eds), *Cognition and categorization*, Lawrence Erlbaum, pp. 27-48. Reprinted in E. Margolis and S. Laurence [1999a].

- [1999], "Reclaiming Concepts", in R. Nunez and W. J. Freeman (eds.), *Reclaiming cognition: The Primacy of Action, Intention and Emotion*, Imprint Academic, pp. 61-77.

Rosch, Eleanor and Mervis, Carolyn [1975], "Family Resemblances: Studies in the Internal Structure of Categories", *Cognitive Psychology* 8, pp. 382-439. Reprinted in M. DePaul and W. Ramsey [1998].

Rosen, Gideon [2002], "A Study in Modal Deviance", in T. Szabó Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*, Oxford University Press, pp. 283-308.

Russell, Gillian [2008], *Truth in Virtue of Meaning*, Oxford University Press.

Sainsbury, Mark [1995], "Vagueness, Ignorance and Margin for Error", *British Journal for the Philosophy of Science* 46, pp. 589-601.

- [1997], "Easy Possibilities", *Philosophy and Phenomenological Research* 57, pp. 907-19.

Schroeter, Laura [2005], "Considering Empty Worlds as Actual", *Australasian Journal of Philosophy* 83, pp. 331-47.

Schulz, Eric, Cokely, Edward T. and Feltz, Adam [forthcoming], "Persistent Bias in Expert Judgements about Free Will and Moral Responsibility: A Test of the Expertise Defense", forthcoming in *Consciousness and Cognition*.

Schwitzgebel, Eric and Cushman, Fiery [forthcoming], "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers", forthcoming in *Mind and Language*.

Shanteau, James [1992], "Competence in Experts: The Role of Task Characteristics", *Organizational Behavior and Human Decision Processes* 53, pp. 252-66.

- [2000], "Why do experts disagree?", in B. Green *et al.* (eds.), *Risk Behaviour and Risk Management in Business Life.* Kluwer Academic Publishers, pp. 186-96.

Shanteau, James, Weiss, David J., Thomas, Rick and Pounds, Julia [2002], "Performance-Based Assessment of Expertise: How to Decide if Someone is an Expert or not", *European Journal of Operations Research* 136, pp. 253-263.

Simon, Herbert and Chase, William G. [1973], "Skill in Chess", *American Scientist* 61, pp. 394-403.

Sorensen, Roy [1992], *Thought Experiments*, Oxford University Press.

Sosa, Ernest [2007a], "Intuitions: their Nature and Epistemic Efficacy", *Grazer Philosophische Studien* 74, pp. 51-67.

- [2007b], *A Virtue Epistemology. Apt Belief and Reflective Knowledge, vol. 1*, Oxford University Press.

- [2007c], "Experimental Philosophy and Philosophical Intuition", *Philosophical Studies* 132, pp. 99-107.

- [2009], "Replies to commentators", *Philosophical Studies* 144, pp. 137-47.

- [2010], "Intuitions and Meaning Divergence", *Philosophical Psychology* 23, pp. 419-26.

Stalnaker, Robert [2001], "On Considering a Possible World as Actual", *Aristotelian Society Supplementary Volume* 75, pp. 141-156.

- [2004], "Assertion Revisited: On the Interpretation of Two-Dimensional Modal Semantics", *Philosophical Studies* 118, pp. 299-322.

- [2011], "The Metaphysical Conception of Analyticity", *Philosophy and Phenomenological Research* 82, pp. 507-14.

Steinberg, Jesse R. [2010], "Dispositions and Subjunctives", *Philosophical Studies* 148, pp. 123-141.

Strawson, Peter [1992], *Analysis and Metaphysics. An Introduction to Philosophy*, Oxford University Press.

Swain, Stacey, Alexander, Joshua and Weinberg, Jonathan [2008], "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp", *Philosophical and Phenomenological Research* 76, pp. 138-65.

Sytsma, Justin and Machery, Edouard [2010], "Two Conceptions of Subjective Experience", *Philosophical Studies* 151, pp. 299-327.

Szabó Gendler, Tamar [1998], "Exceptional Persons: On the Limits of Imaginary Cases", *Journal of Consciousness Studies* 5, pp. 592-610.

- [2002], "Personal Identity and Thought Experiments", *Philosophical Quarterly* 52, pp. 34-54.

Tolhurst, William [1998], "Seeming", *American Philosophical Quarterly* 35, pp. 293-301.

Unger, Peter [1968], "An Analysis of Factual Knowledge", *Journal of Philosophy* 65, pp. 157-70.

- [1984], *Philosophical Relativity*, Oxford University Press.

van Inwagen, Peter [1997], "Materialism and the Psychological-Continuity Account of Personal Identity", *Philosophical Perspectives* 11, *Mind, Causation and World*, pp. 305-319.

Warfield, Ted [1992], "Privileged Self-Knowledge and Externalism are Compatible", *Analysis* 52, pp. 232-37.

- [2005], "Knowledge from Falsehood", *Philosophical Perspectives* 19, pp. 405–16.

Weatherson, Brian [2003], "What good are counterexamples?", *Philosophical Studies* 115, pp. 1-31.

- [2004], "Luminous Margins", *Australasian Journal of Philosophy* 83, pp. 373-83.

Wedgewood, Ralph [2006], "The Normative Force of Reasoning", *Nous* 40, pp. 660-86.

Weinberg, Jonathan [2007], "How to Challenge Intuitions Empirically Without Risking Skepticism", *Midwest Studies in Philosophy* 31, pp. 318–43.

- [2009], "On Doing Better, Experimental Style", *Philosophical Studies* 145, pp. 455-64.

Weinberg, Jonathan, Nichols, Shaun and Stich, Stephen [2001], "Normativity and Epistemic Intuitions", *Philosophical Topics*, pp. 429-60.

Weiss, David J and Shanteau, James [2003], "Empirical Assessment of Expertise", *Human Factors* 45, pp. 104–116.

- [2004], "The Vice of Consensus and the Virtue of Consistency", in C. Smith, J. Shanteau and P. Johnson (eds.), *Psychological explorations of competent decision making*, Cambridge University Press, pp. 226-40.

Weiss, David J., Brennan, Kristin, Thomas, Rick, Kirlik, Alex and Miller, Sarah M. [2009], "Criteria for Performance Evaluation", *Judgment and Decision Making* 4, pp. 164-74.

White, Roger [2006], "Problems for Dogmatism", *Philosophical Studies* 131, pp. 525-57.

Williams, Bernard [1970], "The Self and the Future", *Philosophical Review* 79, pp. 161-80.

- [1973], "Deciding to believe", in B. Williams (ed.), *Problems of the Self*, Cambridge University Press, pp. 136-51.

Williamson, Timothy [2004], "Philosophical 'Intuitions' and Scepticism about Judgement", *Dialectica* 58, pp. 109-153.

- [2005], "Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking", *Proceedings of the Aristotelian Society* 105, pp. 1-23.

- [2007], *The Philosophy of Philosophy*, Blackwell Publishing.

- [2009], "Replies to Ichikawa, Martin and Weinberg", *Philosophical Studies* 145, pp. 465-76.

- [2011a], "Reply to Boghossian" *Philosophy and Phenomenological Research* 82, pp. 498-506.

- [2011b], "Reply to Stalnaker" *Philosophy and Phenomenological Research* 82, pp. 515-523.

Wright, Crispin [2004a], "Warrant for nothing (and foundations for free)?", *The Aristotelian Society Supplementary Volume* 78, pp. 167-212.

- [2004b], "Intuition, Entitlement, and the Epistemology of Logical Laws", *Dialectica* 58, pp. 155-75.

- [2007], "The Perils of Dogmatism", in S. Nuccetelli and G. Seay (eds.), *Themes From G.E. Moore. New Essays in Epistemology and Ethics*, Oxford University Press, pp. 25-48.